A Non-parametric Procedure to Estimate a Linear Discriminant Function with an Application to Credit Scoring

Raquel Voorduin

Thesis submitted for the degree of Doctor of Philosophy

Department of Statistics

University of Warwick

Coventry, CV4 7AL

June, 2004

Acknowledgements

I would like to thank Professor John B. Copas for being my supervisor. He guided me with great patience and profound wisdom, always making invaluable suggestions to my work. He also provided much needed comfort and support during difficult times. Without his motivation and guidance, this work could not have been what it is. To him goes my deepest gratitude.

I would also like to thank Professor Henry P. Wynn for his help and support as co-supervisor to my work. His comments and suggestions were always helpful and motivating.

To all the staff in the Statistics Department at Warwick my gratitude for always providing help when needed. In particular, I would like to thank Professor Jim Q. Smith, Dr. Jane Hutton, Professor Wilfrid S. Kendall, Dr. Elke Thönnes, and Dr. Roberto Puch-Solís. Thank you also to Mrs. Paula J. Matthews for her invaluable help and patience during these years.

I would also like to thank my research fellows and student companions at Warwick. Not only did they provide a welcoming forum for discussion, but they also helped create a friendly atmosphere for work. In particular, many thanks to Miss Beatriz Peñaloza Nyssen for her help with the copying and binding of the thesis.

Finally, I would like to thank the University of Warwick and the United Kingdom government for supporting me with a special research studentship. These thanks also go to Banco de México and CONACYT for their financial support during this time.

Dedication

I would like to dedicate this work to my husband, Marcos. Without his support I would have never completed this journey. To him goes all my love and respect.

To my dear children, Juan Pablo, and María Fernanda who are the most wonderful gifts life has given me.

To my dearest parents, Mauricio and Laura, who have always guided and supported me with love and wisdom.

To my wonderful siblings Mauricio, Laurita, Karen, and Stephanie and to their just as wonderful partners Gloria, Alonso, and Jorge.

To my beautiful nephew Alonso and beautiful niece Alex.

Contents

Chapter 1

Introduction

The present work studies the application of two group discriminant analysis in the field of credit scoring. The view here given provides a completely different approach to how this problem is usually targeted. Credit scoring is widely used among financial institutions and is performed in a number of ways, depending on a wide range of factors, which include available information, support data bases, and informatic resources. Since each financial institution has its own methods of measuring risk, the ways in which an applicant is evaluated for the concession of credit for a particular product are at least as many as credit concessioners. However, there exist certain standard procedures for different products. For example, in the credit card business, when databases containing applicant information are available, usually credit score cards are constructed. These score cards provide an aid to qualify the applicant and decide if he or she represents a high risk for the institution or, on the contrary, a good investment. Score cards are generally used in

conjunction with other criteria, such as the institution's own policies.

In building score cards, generally parametric regression based procedures are used, where the assumption of an underlying model generating the data has to be made. Another aspect is that, in general, score cards are built taking into consideration only the probability that a particular applicant will not default.

In this thesis, the objective will be to present a method of calculating a risk score that does not depend on the actual process generating the data and that takes into account the costs and profits related to accepting a particular applicant. The ultimate objective of the financial institution should be to maximise profit and this view is a fundamental part of the procedure presented here.

1.1 Classification Techniques for Credit Scoring

A range of parametric and non-parametric techniques have been developed and used to model binary data. For the case of credit information, such methods include

- Classic linear discriminant analysis (see Hand,1992).
- Parametric regression techniques (both linear and logistic) which may be found in the class of regression models introduced by McCullagh (1980).
- Nonparametric regression procedures such as kernel methods (see Terrell and Scott, 1992), k-nearest neighbours (see Loftsgarden and Queensberry, 1965) and spline smoothing (see Stone, 1977). These methods use local averaging in the predictor

space to estimate the probability of success

- Decision trees and decision graphs (see Breiman et al., 1984). These procedures include CART (Classification and Regression Trees) and CHAID (chi-squared automatic interaction detection), which are often called "recursive partitioning" since they segment the population into exhaustive, mutually exclusive groups, and
- Neural networks (see Beale and Jackson, 1999)

among others.

It is the experience of the author that credit information is commonly modelled following two main levels. The first level uses procedures such as decision trees to segment the population into homogeneous groups with respect to the response variable (probability of success, for example). The population is not always previously segmented and therefore, only one level of modelling is required.

The second level traditionally uses both linear and logistic parametric regression techniques to model the probability of success in each segment of the population obtained in the first level. Neural networks are also used for second level analysis. They are mainly applied when there is a vast amount of information indicating behavioural patterns, such as credit card transactions, and are used to detect the probability of a transaction having a certain characteristic (being a fraudulent one, for example).

The work in this thesis concentrates on the second level of analysis mentioned above. However, the approach is completely different from traditional procedures in the sense that the focus is placed on the ultimate utility that an "acceptance" (accepted credit

applicant) will provide. The balance between the profits of the successes and the losses of the failures and how this balance can be maximised is the objective of the analysis.

Chapter 2

Utility Function

2.1 Motivation

In the field of two group discriminant analysis the aim is usually to find a score s that will discriminate between both populations. Ideally, large values of this score will correspond to units of one of the populations and small values of the score to units of the other one. If the two populations are labelled as $y = 1$ and $y = 0$ respectively and for each unit there exists a set of covariates x, one can think of the following situation.

Let $y = 1$ denote success and suppose that high values of s correspond to high probabilities of success (the opposite case is completely analogous). Then a simple utility function is given by:

$$U = c_1 \Pr(y = 1 \mid s > k) \Pr(s > k) - c_2 \Pr(y = 0 \mid s > k) \Pr(s > k) \qquad (2.1)$$

where c_1 and c_2 are two known positive costs. In the field of credit scoring, where a "good" customer is defined as one who repays his or her loan, this can be seen as a loss function that adds a positive quantity c_1 (profit) for every good customer that was accepted and subtracts a positive quantity c_2 (loss) for every accepted "bad" customer. The rejected individuals (those having $s \leq k$) are not considered regardless of being potentially good or bad. The business will usually have ways to evaluate these profits and losses (c_1 and c_2) and this issue is in itself worth exploring. Generally, these "costs" will depend on different variables, which may include the covariates obtained at the time of application. Particular features of the credit product, including the credit line given, along with the costs of doing business, interest rates, and the macro-economic environment will affect the overall cost of a customer. Usually, banks calculate a profitability index for new accounts, a number which indicates the amount of money they expect to win (or lose) with that particular customer and which involves all these variables. Since the bank cannot know in advance the profit a customer will generate, the profitability index is, in fact, a model in itself. Therefore, taking c_1 and c_2 as constants represents a simplification of the real problem. In §6.2 some insight around the idea of having the cost depend on covariates will be given.

The utility function above will keep its basic form if it is re-scaled by making $c_1 + c_2 = 1$. In the following expression, if $c = c_2$, then

$$U = \Pr(s > k)\,[\Pr(y = 1 \mid s > k) - c]$$

Here k is a threshold that indicates inclusion. The units having $s \leq k$ will neither have a positive nor a negative effect on the utility. Now suppose s is given by a linear combination of the covariates, say $s = \beta^T x$, and that f_x is the density of x and $p_x = \Pr(y = 1 \mid x)$, the probability of success. The probability of inclusion (acceptance in credit scoring) is

$$\Pr(s > k) = \Pr(\beta^T x > k) = \int_{\beta^T x > k} f_x dx$$

In the above expression x does not include an intercept term since this is the same thing as having $k \neq 0$ and this would produce a redundancy of the parameters in $\beta^T x > k$. In fact, k is also redundant, since the inclusion criteria is unaffected if $(\beta/k)^T x > 1$ is used. So, in order to avoid overparametrisation, in the future $\beta^T x > 1$ (covariates x lacking an intercept term) will be used to indicate inclusion. Then, the utility function may be expressed as:

$$U_\beta = \int_{\beta^T x > 1} (p_x - c) f_x dx \tag{2.2}$$

where the subindex in U_β indicates utility is a function of the parameter β of the score. Given p_x and c, we aim to find β such that U_β is maximised.

2.2 Overview

Literature exploring discriminant analysis and logistic regression models is wide and varied. This work is centred on the particular application of a discriminating technique to the

credit scoring field. Linear discriminant analysis, as introduced by Fisher (Fisher, 1936) is a procedure based on the normal distribution theory. Suppose a sample size n_1 and a sample size n_2 are taken from each of two populations, where the data matrix, denoted by X_i contains the measurement of r covariates for n_i individuals, i.e. X_i is order $n_i \times r$ for $i = 1, 2$. Furthermore, X_i is taken to be a random sample, size n_i, of observations from the distribution $N_r(\mu_i, \Sigma)$, where a same covariance matrix is assumed for both populations. Then (see Chatfield (1995))

$$\bar{X}_1 - \bar{X}_2 \sim N_r\left(\mu_1 - \mu_2, \left[\frac{1}{n_1} + \frac{1}{n_2}\right]\Sigma\right)$$

An estimate for Σ is given by the pooled within-groups sample variance-covariance matrix S

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

where S_1 and S_2 represent the sample variance-covariance matrix for each population. The linear score $s = \beta^T X$ where β is the solution to

$$S\beta = (\bar{x}_1 - \bar{x}_2)$$

is called the linear discriminant function and was first proposed by Fisher.

An alternative approach is discussed in Eguchi and Copas (2002), where discrimination is based upon minimising a risk function. This function depends on a score based on covariates x, say $s = \beta^T x$, and will increasingly penalise high values s when $y = 0$

and low values of it when $y = 1$. Such a risk function is written as

$$D(\beta) = E\{-yU(s) + (1-y)V(s)\} \tag{2.3}$$

where $U(s)$ and $V(s)$ are two monotonically increasing functions of the score (s). The functions U and V are then assumed to be logistically consistent. This means that, if the logistic model happens to be correct, that is, if the probability of success given x is

$$p(x) = \Pr(Y = 1|x) = p_L(s) = \frac{e^s}{1+e^s}$$

where $s = \beta^T x$, then $D(\beta)$ will be minimised at the true value of β. Eguchi and Copas show that this assumption leads to

$$\frac{\partial D(\beta)}{\partial \beta} = -E[xW(s)\{p(x) - p_L(s)\}] \tag{2.4}$$

where

$$W(u) = w(u)(1+e^u)$$

and

$$w(u) = \frac{\partial U(u)}{\partial u} = e^{-u}\frac{\partial V(u)}{\partial u}$$

and where w can be thought of as a weight function. So U and V are both functions of the weight w. Also, (2.4) is a function of the unknown parameter β since $p_L(s)$ is a

function of it.

In their paper, Eguchi and Copas describe a credit scoring situation where applicants are accepted only if $s \geq u$, the cut-off. The aim is to maximise expected profit given by

$$a_1 \Pr(s \geq u | y = 1) \Pr(Y = 1) - a_0 \Pr(s \geq u | y = 0) \Pr(Y = 0) \tag{2.5}$$

where a_1 is the profit when $y = 1$ and a_0 the loss when $y = 0$ for each accepted applicant. In Example 2, §2.4 of the paper "$w(u) = \delta(u - u_0)$. Here the weight function is the Dirac delta function at $u = u_0$ for some fixed value u_0. In this case

$$U(u) = H(u - u_0), \quad V(u) = e^{u_0} H(u - u_0),$$

where H is the Heaviside function $H(u) = 1$ if $u \geq 0$ and zero otherwise." If u_0 is given by

$$u_0 = \log\left(\frac{a_0}{a_1}\right)$$

where a_0 represents the cost of accepting an individual from population $y = 0$ and a_1 represents the profit of accepting an individual from population $y = 1$, this corresponds to the credit scoring example. Minimising $D(\beta)$ for the credit scoring example in (2.5) is equivalent to solving (2.4) where $w(u) = \delta(u - u_0)$ and u_0 as above.

In the discussion of the paper, the logistic consistency of the functions U and V is an essential part of the derivation of discriminant functions. The estimation procedure followed uses smoothed versions of the functions U and V (obtained using a kernel method).

Weighted logistic regression is then used to estimate the parameter β, where the weights are given by $W(\beta^T x)$. This procedure is followed iteratively since the weights depend on the unknown parameter, using the unweighted logistic regression estimate of β as starting point. The estimation procedure followed in this thesis is somewhat different, since it is centred around estimating the derivative of (2.5) directly, using a kernel method and solving this for β. The optimal bandwidth in this case is obtained analytically. This is different from the procedure followed in the paper by Eguchi and Copas, where a near-logistic setting (allowing for mis-specification of the model) arises the necessity to compromise between a parameter estimate which is best for variance (in the logistic setting) and one that is best for bias (near-logistic setting). The choice of optimal bandwidth is then replaced by cross-validation on a mixture parameter that balances this compromise in the choice of a weight function.

The above ideas from the paper along with the derivation of sampling properties of β, crossvalidation risk, and the discussion of the method for different sampling schemes were all fundamental to the development of similar ideas in this thesis. In particular, the case study presented here, being a cohort study in the sense that sampling is conditional on x, takes into consideration weights which are the inverses of the probabilities of selection, as stated in the paper. Cross-validation corrections presented are mostly based on the ideas discussed in this paper as well.

2.3 Layout of thesis

The thesis is organised in six main sections. These sections are presented in the natural order of development of the thesis. However, some of them have parallel positions in terms of the sequence of derivation.

First of all, the theory for non-parametric estimation is given for the general case dealing with one or more covariates. Here, the idea of maximising a utility function is explored and parameters are estimated under this maximisation. Expressions for the bias and variance of the estimates of the parameters are obtained and through these, the optimal window width. Special cases are explored that may lead to simplification of the expressions. The next section provides an insight into cross-validated approximations for the estimated utility and the empirical assessment of the estimates through two examples obtained by simulation processes. Using only one covariate may lead to a different, more direct, parametrisation. This is explored in §5, where the link to quantal bioassay, a direct application, is recognised. Again, two examples obtained by simulation processes are analysed. The next two chapters deal with generalisations of different kinds. In §6 a generalisation of the non-parametric formulation involving different weights in the objective function is discussed. Also, a more general view of the cost function involved in utility is explored. Chapter 7 talks about a special relocation of the covariates in the multivariate case. Finally, a case study is presented showing an application to most of the results and giving some insight into the special features of credit card application information.

An additional point, the title of the thesis refers to a non-parametric procedure. This is because the estimation process is essentially non-parametric. However, a linear score is fundamental to the basic idea. In this sense, the term non-parametric could perhaps be modified to semi-parametric.

Chapter 3

Theory for Non-parametric Estimation

3.1 Maximisation of utility function

The objective is to find a value of β that will maximise utility given by U_β in (2.2). Therefore the value of $\hat{\beta}$ looked for is the root of $U'_\beta = 0$. An estimate for β may be found using a Newton type approximation. To be able to do this, it is required that $U'_\beta = 0$ complies with standard regularity conditions such as being continuous and differentiable.So,

$$\hat{\beta} \simeq \beta_0 - \left(\frac{\partial^2 U_\beta}{\partial \beta^T \partial \beta}(\beta_0)\right)^{-1} \left(\frac{\partial U_\beta}{\partial \beta}(\beta_0)\right)$$

where β_0 is an initial value. An iterative procedure is followed using as β_0 the estimate for β obtained in the previous iteration. The process is stopped when the difference

between $\hat{\beta}$ and β_0 is relatively small, which is the same as having $\frac{\partial U_n}{\partial \beta}(\beta_0) \simeq \frac{\partial U_n}{\partial \beta}\left(\hat{\beta}\right)$ very close to zero. To obtain a derivative of (2.2) with respect to β, consider Figure 3.1.

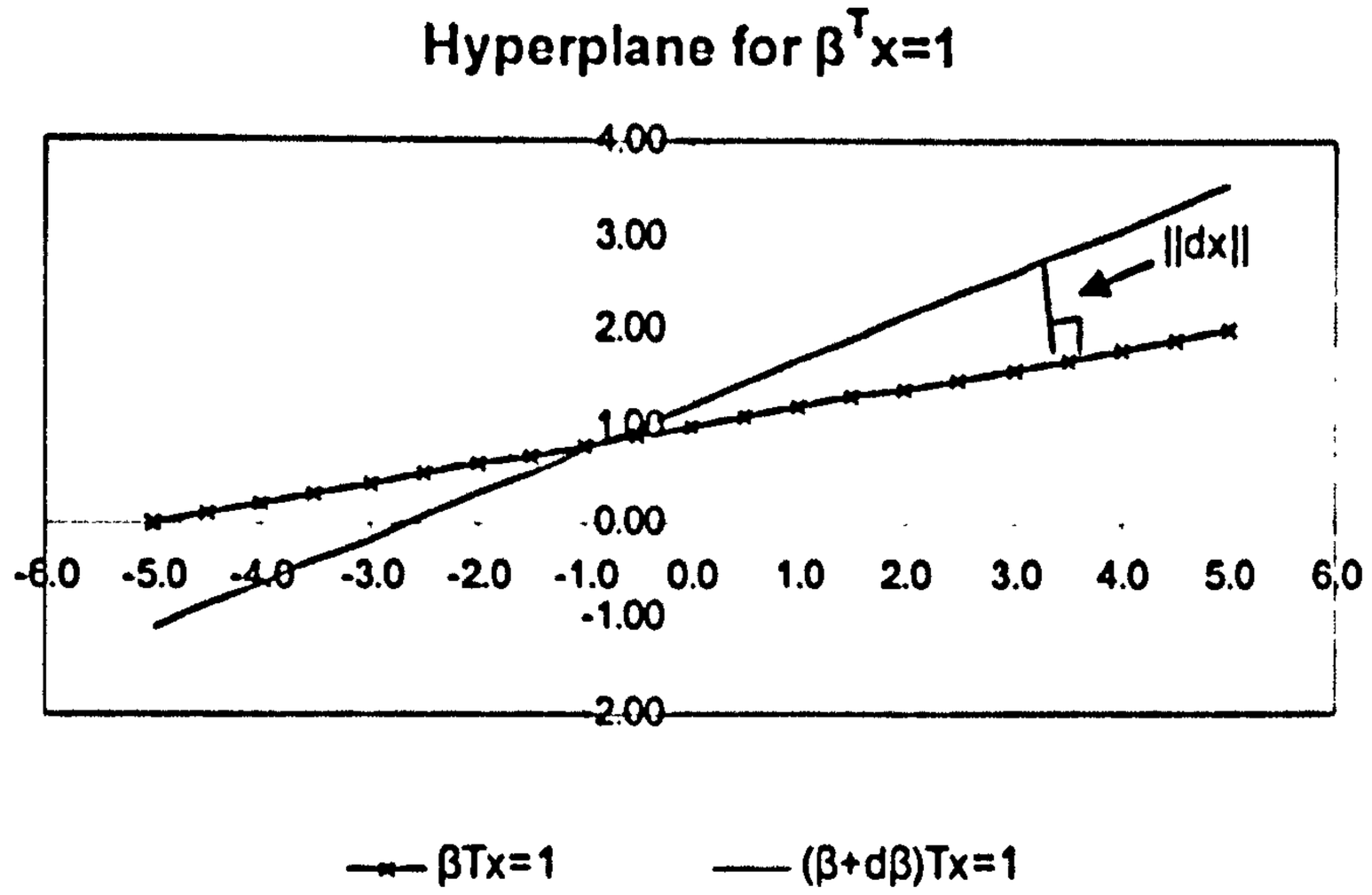


Figure 3.1

For simplicity, we will first consider the derivative with respect to β_1, the parameter of the first covariate. Take

$$(d\beta)^T = (d\beta_1, 0, 0, ..., 0)$$

and $v_x = (p_x - c)\, f_x$ Then

$$\frac{\partial U_\beta}{\partial \beta_1} = \lim_{d\beta_1 \to 0} \frac{\int_{(\beta+d\beta)^T x>1} v_x dx - \int_{\beta^T x>1} v_x dx}{d\beta_1}$$

$$= \lim_{d\beta_1 \to 0} \frac{\int_{\beta^T x=1} \|dx\| v_x dx}{d\beta_1}$$

where $\|dx\|$ represents the height of the distance between $\beta^T x + d\beta^T x = 1$ and $\beta^T x = 1$.

We have the simultaneous equations

$$(\beta + d\beta)^T (x + dx) = \beta^T x + \beta^T dx + d\beta_1 x_1 + d\beta_1 dx_1 = 1$$

$$\beta^T x = 1$$

These equations are subtracted and then, since $d\beta_1 dx_1$ is small, $\beta^T dx = -d\beta_1 x_1$. Also, since dx runs in the same direction as β then

$$dx = l\beta$$

where l is a scalar. Then

$$\beta^T (l\beta) = -d\beta_1 x_1$$

$$l = -\frac{d\beta_1 x_1}{\beta^T \beta}$$

Therefore

$$dx = -\frac{\beta}{\beta^T\beta} d\beta_1 x_1$$

and

$$\|dx\| = \frac{\|\beta\|}{\beta^T\beta} d\beta_1 x_1 = \frac{1}{\sqrt{\beta^T\beta}} d\beta_1 x_1$$

so

$$\begin{aligned} \frac{\partial U_\beta}{\partial \beta_1} &= \lim_{d\beta_1 \to 0} \frac{\int_{\beta^T x=1} \|dx\| v_x dx}{d\beta_1} = \lim_{d\beta_1 \to 0} \frac{1}{\sqrt{\beta^T\beta}} \frac{\int_{\beta^T x=1} d\beta_1 x_1 v_x dx}{d\beta_1} \\ &= \frac{1}{\sqrt{\beta^T\beta}} \int_{\beta^T x=1} x_1 v_x dx \end{aligned}$$

And therefore

$$U'_\beta = \frac{\partial U_\beta}{\partial \beta} = \frac{1}{\sqrt{\beta^T\beta}} \int_{\beta^T x=1} x\,(p_x - c)\, f_x dx \tag{3.1}$$

To obtain the second derivative of utility with respect to β, let $w_x = x v_x = x\,(p_x - c)\, f_x$.

Following the same idea, the difference

$$\int_{(\beta+d\beta)^T x=1} w_x dx - \int_{\beta^T x=1} w_x dx$$

is required. By Taylor,

$$w_{x+dx} \simeq w_x + w'_x dx$$

where $dx = -\frac{\beta x^T d\beta}{\beta^T \beta}$ may be found in a similar way as before, solving the equation system

$$\begin{aligned} (\beta + d\beta)^T (x + dx) &= 1 \\ \beta^T x &= 1 \end{aligned}$$

where now $(d\beta)^T = (d\beta_1, d\beta_2, ..., d\beta_r)$. Therefore,

$$\begin{aligned} \int_{(\beta+d\beta)^T x=1} w_x dx &\simeq \int_{\beta^T x=1} w_{x+dx} dx \\ &\simeq \int_{\beta^T x=1} w_x dx - \int_{\beta^T x=1} w'_x \frac{\beta x^T d\beta}{\beta^T \beta} dx. \end{aligned}$$

So,

$$\frac{\partial}{\partial \beta} \int_{\beta^T x=1} w_x dx = - \int_{\beta^T x=1} w'_x \frac{\beta x^T d\beta}{\beta^T \beta} dx$$

where

$$w'_x|_{ij} = \frac{\partial}{\partial x_j} v_x x_i = \begin{cases} x_i \frac{\partial}{\partial x_j} v_x & i \neq j \\ v_x + x_i \frac{\partial}{\partial x_i} v_x & i = j \end{cases}$$

and $v_x = (p_x - c) f_x$ as before. Therefore,

$$w'_x = v_x I + x v'^T_x = v_x \left\{ I + x \left(\frac{\partial}{\partial x} \log v_x \right)^T \right\}$$

and

$$\frac{\partial}{\partial \beta} \int_{\beta^T x=1} x v_x dx = -\frac{1}{\beta^T \beta} \int_{\beta^T x=1} \left\{ I + x \left(\frac{\partial}{\partial x} \log v_x \right)^T \right\} v_x \beta x^T dx.$$

Then

$$U''_\beta = \frac{\partial^2 U_\beta}{\partial\beta\partial\beta^T} = -\frac{1}{(\beta^T\beta)^{3/2}} \times \int_{\beta^T x=1} \left[I + x\left[\frac{f'_x(p_x-c)+p'_x f_x}{(p_x-c)f_x}\right]^T\right](p_x-c)f_x\beta x^T dx \tag{3.2}$$

Maximum utility will be obtained when $U'_\beta = \frac{\partial U_\beta}{\partial\beta} = 0$. If U'_β is divided by $\int_{\beta^T x=1} f_x dx$ then this is equivalent to saying that

$$\frac{U'_\beta}{\int_{\beta^T x=1} f_x dx} = E\left[x(p_x-c) \mid \beta^T x = 1\right] = 0$$

Multiplying by β both sides of the equation gives

$$E\left[(p_x-c) \mid \beta^T x = 1\right] = 0 \tag{3.3}$$

The expectation is taken over all individuals belonging to the hyperplane $\beta^T x = 1$. If p_x is a function of $\beta^T x$ then the above equation would imply solving $p_x = c$ when $\beta^T x = 1$. In this case, the second derivative of utility would reduce to

$$U''_{\beta 1} = -\frac{p'^T\beta}{(\beta^T\beta)^{3/2}} \int_{\beta^T x=1} xx^T f_x dx$$

where p' represents the derivative of p_x with respect to x, evaluated at $\beta^T x = 1$. This happens because at the solution, when $U'_\beta = 0$ and $p_x = c$ the terms involving I and f'_x

will vanish in (3.2).

A further assumption to simplify the hessian in U''_β can be made. Suppose p_x is given by a specific model. For example, suppose p_x is logistic, then $\log\left(\frac{p_x}{1-p_x}\right) = \gamma_0 + \gamma^T x$, a value of β can be found when $p_x = c$ in order to have $\beta^T x = 1$. Taking $\beta = \gamma / \left[\log\left(\frac{c}{1-c}\right) - \gamma_0\right]$ is the value of β such that $\beta^T x = 1$ if p_x is given by the proposed logistic model. In this case the derivative of p_x with respect to x can be obtained, $p'_x = \gamma^T p_x (1 - p_x)$. Then

$$U''_{\beta l} = -\frac{\beta^T \gamma c (1-c)}{(\beta^T \beta)^{3/2}} \int_{\beta^T x = 1} x x^T f_x dx \tag{3.4}$$

So $U''_{\beta l}$ is the expression for the second derivative of utility when p_x is logistic.

In practice, estimates for utility and its first and second derivatives will be needed. In this situation, virtually no data will fall exactly on the hyperplane given by $\beta^T x = 1$. The following expressions consider a rectangular window around $\beta^T x = 1$ for estimation purposes. When working with a sample of size n, if response is given by y_i for each unit, estimates for the above expressions may be obtained by

$$\hat{U}_\beta = \frac{1}{n} \sum_{\beta^T x > 1} (y_i - c) \tag{3.5}$$

$$\hat{U}'_\beta = \frac{1}{n} \frac{1}{\sum_D 1} \frac{1}{\sqrt{\beta^T \beta}} \sum_D x_i (y_i - c) \tag{3.6}$$

Here, D is the subset including all individuals in a vicinity of $\beta^T x = 1$. That is, D

contains all individuals with a value for $\beta^T x$ falling within $(1-h, 1+h)$ where h is taken to be relatively small. Here $\hat{U}_\beta$ and $\hat{U}'_\beta$ are normalised using a factor of $\frac{1}{n}$, indicating that utility is per individual in the sample. An approximation to U''_β is given, using the logistic assumption. This assumption is used in order to simplify the hessian of the second derivative and will be useful to set up the numerical method. However, the estimates themselves will not be affected. In fact, any other model assumption for p_x could be helpful in order to approximate this matrix. We define

$$\hat{U}''_{\beta l} = -\frac{1}{n}\frac{1}{\sum_D 1}\frac{\beta^T \hat{\gamma} c(1-c)}{(\beta^T \beta)^{3/2}} \sum_D x_i x_i^T \tag{3.7}$$

where $\hat{\gamma}$ is the logistic estimate for γ. One way to look at $\hat{U}'_\beta$ and $\hat{U}''_{\beta l}$ is the following:

$$\hat{U}'_\beta = \frac{1}{n}\frac{1}{\sum_{i=1}^n w_i}\frac{1}{\sqrt{\beta^T \beta}} \sum_{i=1}^n w_i x_i (y_i - c) \tag{3.8}$$

$$\hat{U}''_{\beta l} = -\frac{1}{n}\frac{1}{\sum_{i=1}^n w_i}\frac{\beta^T \hat{\gamma} c(1-c)}{(\beta^T \beta)^{3/2}} \sum_{i=1}^n w_i x_i x_i^T \tag{3.9}$$

$$w_i = \begin{cases} 1 \text{ if } 1-h \leq \beta^T x \leq 1+h \\ 0 \text{ otherwise} \end{cases}$$

There exist a number of options for the choice of w_i, not just a rectangular set. Also, a factor of $\frac{1}{\sum_{i=1}^n w_i}$ so that the weights used add up to one is included. The possibility that we explore in the following section is to consider a Gaussian kernel for the weight w_i. So $w_i = \phi\left(\frac{\beta^T x_i - 1}{h}\right)$ where ϕ represents the standard normal density and the width of the window depends on h again.

3.1.1 Special Cases

In rare cases, there will be situations where the optimal solution cannot be found. This may happen, for example if c_1 or c_2 are equal to zero in (2.1). If $c_1 = 0$, that is, if the profit of a success is zero, then we would logically reject everybody. Then utility would be zero, which is the maximum in this context. On the contrary, if $c_2 = 0$, this would imply the loss for every failure is zero and we should accept everybody in order to maximise utility.

Another extreme case may arise if every individual has same values for every covariate in x. In this sense, it would not be possible to find any difference between individuals and discriminate between the "good" and the "bad".

These are only two examples where the maximum utility will not be contained within the span of the data, but one could possibly think of similar situations. The non-parametric estimation procedure followed in this work assumes that an optimal solution exists within the span of the data. Therefore, it is always important to study the nature of the utility structure in order to detect situations such as these, where different approaches should be followed.

3.2 Non-parametric estimation

The objective of the present section is to formulate the problem using a Gaussian kernel for the estimation of β. The aim is to find the window width h that will be best for

expected utility. The asymptotic properties of the bias and variance of the estimate of β and the fact that in this application n will be usually very large are taken into consideration.

Taking (3.8) with weights $w_i = \phi\left(\frac{\beta^T x_i - 1}{h}\right)$ and letting h be the window width

$$A_\beta = \frac{1}{nh}\sum_{i=1}^{n} x_i (y_i - c)\,\phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{3.10}$$

$$A'_\beta = -\frac{1}{nh}\sum_{i=1}^{n} x_i x_i^T (y_i - c)\left(\frac{\beta^T x_i - 1}{h^2}\right)\phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{3.11}$$

Here, the factor $\frac{1}{\sqrt{\beta^T\beta}}$ is eliminated because the root of A_β will be the same if this constant is included or not. Also, an estimate for β would be given by $\hat{\beta}_h \simeq \beta - \left[A'_\beta\right]^{-1} A_\beta$, the subindex h in $\hat{\beta}_h$ indicates that it is an estimate of β that depends on the value of h, the window width of the kernel function used. However, to make the estimation process more stable, the expectation of A'_β is used, using instead $\hat{\beta}_h \simeq \beta - \left[E\left(A'_\beta\right)\right]^{-1} A_\beta$ where

$$E_y\left(A'_\beta\right) = -\frac{1}{nh}\sum_{i=1}^{n} x_i x_i^T (p_{x_i} - c)\left(\frac{\beta^T x_i - 1}{h^2}\right)\phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{3.12}$$

where the expectation is taken over y. Expanding p_x around $\beta^T x = 1$ and keeping only first order terms so that $p_x \simeq p + \left(\beta^T x - 1\right) p'$ gives

$$E_y\left(A'_\beta\right) \simeq -\frac{1}{nh}\sum_{i=1}^{n} x_i x_i^T p'\left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{3.13}$$

If a logistic model such as the one used in the previous section is assumed for p_x then

$$E_y\left(A'_\beta\right) \simeq -\frac{bc(1-c)}{nh}\sum_{i=1}^{n} x_i x_i^T \left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{3.14}$$

where $b = \left[\log\left(\frac{c}{1-c}\right) - \gamma_0\right]$. Here, the root of (3.3) is given by $\beta = \gamma / \left[\log\left(\frac{c}{1-c}\right) - \gamma_0\right]$ so $p_x = \frac{\exp\left(\gamma_0 + b\beta^T x\right)}{1+\exp\left(\gamma_0 + b\beta^T x\right)}$ and its derivative with respect to $\beta^T x$ is $p'_x = bp_x(1-p_x)$, which evaluated at $\beta^T x = 1$ gives $p' = bc(1-c)$.

3.3 Approximations for bias and variance of $\hat{\beta}_h$

To obtain approximations for the bias and variance of $\hat{\beta}_h$, a set of functions of the score, involving expectations of functions of x and p_x will be defined. This will help simplify notation in the later calculations. Let

$$M_s = E_x\left[x\left(p_x - c\right) \mid \beta^T x = s\right] \tag{3.15}$$

$$N_s = E_x\left[xx^T\left(p_x - c\right) \mid \beta^T x = s\right] \tag{3.16}$$

$$W_s = E_x\left[xx^T p_x\left(1-p_x\right) \mid \beta^T x = s\right] \tag{3.17}$$

where the subindex indicates the expectation is over x. The true value of the parameter β, is given by $M = M_1 = 0$ (in every case, the absence of argument will indicate evaluation of the function at $s = \beta^T x = 1$). Suppose that $\hat{\beta}_h$ is obtained using $\hat{\beta}_h \simeq \beta - \left[E\left(A'_\beta\right)\right]^{-1} A_\beta$ where A_β and $E\left(A'_\beta\right)$ are as defined in the previous section. Then the bias and variance

of $\hat{\beta}_h$ are

$$\begin{aligned} Bias\left(\hat{\beta}_h\right) &= E_y\left[\hat{\beta}_h - \beta\right] \simeq -\left[E_y\left(A'_\beta\right)\right]^{-1} E_y\left(A_\beta\right) \\ Var\left(\hat{\beta}_h\right) &\simeq \left[E_y\left(A'_\beta\right)\right]^{-1} Var_y\left(A_\beta\right)\left[E_y\left(A'_\beta\right)\right]^{-1} \end{aligned}$$

Here, the subindex indicates expectation is over y. Using (3.10), (3.12), and the above expressions for M_s, N_s, and W_s the expectations become

$$E_y\left(A_\beta\right) = \frac{1}{h} E_s\left[M_s \phi\left(\frac{s-1}{h}\right)\right] \tag{3.18}$$

$$E_y\left(A'_\beta\right) = -\frac{1}{h^2} E_s\left[N_s\left(\frac{s-1}{h}\right)\phi\left(\frac{s-1}{h}\right)\right] \tag{3.19}$$

$$Var_y\left(A_\beta\right) = \frac{1}{nh^2} E_s\left[W_s \phi^2\left(\frac{s-1}{h}\right)\right] \tag{3.20}$$

Expectations are now taken over the distribution of s with density $g_s = \frac{d}{ds}\Pr\left(\beta^T x < s\right)$ and ϕ represents the standard normal density. That is

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}u^2\right\}. \tag{3.21}$$

Changing variable to $u = \frac{s-1}{h}$ so that $du = \frac{1}{h}ds$ and $s = 1 + hu$ gives

$$E_y\left(A_\beta\right) = \int_{-\infty}^{+\infty} \phi_u M_{1+hu} g_{1+hu} du$$

Expanding each function using a Taylor series, keeping only first order terms will give

$M_{1+hu} \simeq M_1 + M_1' hu + \frac{1}{2} M_1'' h^2 u^2$, since $M_1 = 0$ by definition and $g_{1+hu} \simeq g_1 + g_1' hu$.

Therefore the product of these two functions is approximated as

$$M_{1+hu} g_{1+hu} = M_1 g_1 + M_1' g_1 hu + M_1 g_1' hu + M_1' g_1' h^2 u^2 + \frac{1}{2} M_1'' g_1 h^2 u^2 + \frac{1}{2} M_1'' g_1' h^3 u^3$$

where only the term in u^2 will survive. This happens because on one hand $M_1 = 0$, which leaves out the first term, and on the other, because the expectation of terms having odd powers of u are zero since ϕ_u is standard normal. Since $\int_{-\infty}^{+\infty} \phi_u u^2 du = 1$ then

$$E_y(A_\beta) = \int_{-\infty}^{+\infty} \phi_u M_{1+hu} g_{1+hu} du \simeq h^2 \left[M' g' + \frac{1}{2} M'' g \right] \tag{3.22}$$

An analogous argument is followed in order to obtain the next approximation

$$E_y(A_\beta') = -\frac{1}{h} \int_{-\infty}^{+\infty} u \phi_u N_{1+hu} g_{1+hu} du \simeq -\left[N g' + N' g \right] \tag{3.23}$$

For the approximation of $Var_y(A_\beta)$, the variable change $u = \frac{\sqrt{2}(s-1)}{h}$ so that $du = \frac{\sqrt{2}}{h} ds$ and $s = 1 + \frac{h}{\sqrt{2}} u$ gives

$$Var_y(A_\beta) = \frac{1}{nh\sqrt{4\pi}} \int_{-\infty}^{+\infty} \phi_u W_{1+\frac{h}{\sqrt{2}}u} g_{1+\frac{h}{\sqrt{2}}u} du \simeq \frac{1}{nh\sqrt{4\pi}} W g$$

Here M', g', M'', and N' indicate derivatives with respect to s. Using these expressions

the bias and variance for $\hat{\beta}_h$ become

$$Bias\left(\hat{\beta}_h\right) \simeq -h^2\left[Ng_L + N'\right]^{-1}\left[M'g_L + \frac{1}{2}M''\right] = -h^2\alpha \tag{3.24}$$

$$Var\left(\hat{\beta}_h\right) \simeq \frac{1}{nh\sqrt{4\pi g}}\left[Ng_L + N'\right]^{-1} W' \left[Ng_L + N'\right]^{-1} = \frac{1}{nh}B \tag{3.25}$$

where

$$g_L = \frac{\partial}{\partial s}\log g\,|_{s=1}$$

$$\alpha = \left[Ng_L + N'\right]^{-1}\left[M'g_L + \frac{1}{2}M''\right]$$

and

$$B = \frac{1}{\sqrt{4\pi g}}\left[Ng_L + N'\right]^{-1} W' \left[Ng_L + N'\right]^{-1}$$

3.4 Optimal window width h

Now the objective is to choose $\hat{\beta}_h$ in order to maximise $E\left(U_{\hat{\beta}_h}\right)$. Since an approximation to $U_{\hat{\beta}_h}$ can be given by $U_{\hat{\beta}_h} \simeq U_\beta + \left(\hat{\beta}_h - \beta\right)^T U'_\beta + \frac{1}{2}\left(\hat{\beta}_h - \beta\right)^T U''_\beta \left(\hat{\beta}_h - \beta\right)$ where $U'_\beta = 0$ and U''_β is as in (3.2), then

$$E\left(U_{\hat{\beta}_h}\right) \simeq U_\beta + \frac{1}{2}\left[Bias\left(\hat{\beta}_h\right)^T U''_\beta Bias\left(\hat{\beta}_h\right) + tr\left(Var\left(\hat{\beta}_h\right) \times U''_\beta\right)\right]$$

Maximising with respect to h and substituting the expressions for bias and variance where appropriate gives

$$h = \left\{ \frac{tr\left(BU''_\beta\right)}{4n\alpha^T U''_\beta \alpha} \right\}^{1/5} \propto n^{-1/5} \tag{3.26}$$

It is worth noting that the formula for h is proportional to $n^{-1/5}$ which is of same order of magnitude as the optimal window width that Silverman (1986) finds for density estimation using kernel functions.

3.5 Special Cases

3.5.1 Probability of success as function of the score

When the probability of success is defined entirely by the score, that is, when $p_x = p_s$, then a significant simplification of the above expressions is possible. In this case the following changes apply

$$M_s = (p_s - c)\,\mu_s$$

where $\mu_s = E_x\left[x \mid \beta^T x = s\right]$. It follows that $M = 0 \Leftrightarrow p = c$. Also

$$\begin{aligned} N_s &= (p_s - c)\,V_s \\ N &= 0 \end{aligned}$$

where $V_s = E_x \left[xx^T \mid \beta^T x = s \right]$.

$$W_s = p_s (1 - p_s) V_s$$

$$W = c(1-c) V$$

Here,

$$M'_s = p'_s \mu_s + (p_s - c) \mu'_s$$

$$M' = p' \mu$$

$$M''_s = p''_s \mu_s + 2p'_s \mu'_s + (p_s - c) \mu''_s$$

$$M'' = p'' \mu + 2p' \mu'$$

$$N'_s = p'_s V_s + (p_s - c) V'_s$$

$$N' = p' V$$

From the above expressions α , B , and U''_β change in the following way, thus simplifying the bias and variance of $\hat{\beta}_h$.

$$\alpha = V^{-1} \left[\mu \left(g_L + \frac{1}{2} \frac{p''}{p'} \right) + \mu' \right] = V^{-1} \delta$$

$$B = \frac{c(1-c)}{\sqrt{4\pi} g (p')^2} V^{-1}$$

$$U''_\beta = -\frac{\beta^T \left(\frac{\partial}{\partial x} p \right)}{(\beta^T \beta)} V g$$

where

$$\delta = \mu\left(g_L + \frac{1}{2}\frac{p''}{p'}\right) + \mu'$$

The expression for optimum h is simplified accordingly:

$$h = \left\{\frac{c(1-c)r}{4\sqrt{4\pi}ng(p')^2\delta^T V^{-1}\delta}\right\}^{1/5} \tag{3.27}$$

where r is again the number of covariates.

3.5.2 Probability of success given by a logistic model

In this case, the probability of success is given by a specific known model, for example p_x may be logistic. Then $logit(p_x) = \gamma_0 + \gamma^T x$, and at the solution, $\beta = \gamma/b$, where $b = \log\left(\frac{c}{1-c}\right) - \gamma_0$ so

$$\begin{aligned} p' &= \frac{\partial}{\partial s}p_s\,|_{s=1} = bc(1-c) \\ p'' &= \frac{\partial^2}{\partial s^2}p_s\,|_{s=1} = b^2c(1-c)(1-2c) \end{aligned}$$

If these values for the derivatives of p_s are taken, then δ changes to

$$\delta = \mu\left(g_L + \frac{1}{2}b(1-2c)\right) + \mu'$$

the bias and variance are reduced to

$$Bias\left(\hat{\beta}_M\right) \simeq h^2 V^{-1}\left[\mu\left(g_L + \frac{1}{2}b\left(1-2c\right)\right) + \mu'\right] \tag{3.28}$$

$$Var\left(\hat{\beta}_M\right) \simeq \frac{1}{nh}\frac{1}{\sqrt{4\pi}gb^2c\left(1-c\right)}V^{-1} \tag{3.29}$$

and h can be written as

$$h = \left\{\frac{r}{4\sqrt{4\pi}ngb^2c\left(1-c\right)\delta^T V^{-1}\delta}\right\}^{1/5}. \tag{3.30}$$

3.5.3 Normal distribution for covariates x

Suppose $x \sim N\left(m, \Sigma\right)$. Then $\beta^T x \sim N\left(\beta^T m, \beta^T \Sigma \beta\right)$. If X is a vector formed by x and $\beta^T x$ so that $X = \left[\begin{smallmatrix} x \\ \beta^T x \end{smallmatrix}\right]$ then the distribution of X is given by

$$X \sim N\left[\begin{pmatrix} m \\ \beta^T m \end{pmatrix}; \begin{pmatrix} \Sigma & \Sigma\beta \\ \beta^T\Sigma & \beta^T\Sigma\beta \end{pmatrix}\right]$$

When $\beta^T\Sigma\beta$ is different from zero so that $\left(\beta^T\Sigma\beta\right)^{-1}$ exists, the conditional distribution of x given $\beta^T x = s$ is also multivariate normal with mean and variance given by

$$\mu_s = E\left(x \mid \beta^T x = s\right) = m + \Sigma\beta\left(\frac{1}{\beta^T\Sigma\beta}\right)\left(s - \beta^T m\right) \tag{3.31}$$

$$Var\left(x \mid \beta^T x = s\right) = \Sigma - \frac{\Sigma\beta\beta^T\Sigma}{\beta^T\Sigma\beta} \tag{3.32}$$

Using these expressions, it follows that

$$\mu = E\left[x \mid \beta^T x = 1\right] = m + \Sigma\beta\left(\frac{1}{\beta^T\Sigma\beta}\right)(1 - \beta^T m) \tag{3.33}$$

$$\mu' = \frac{\partial}{\partial s}\mu_s \mid_{s=1} = \frac{\Sigma\beta}{\beta^T\Sigma\beta} \tag{3.31}$$

The expressions for the density of $\beta^T x$ are

$$g_{\beta^T x} = \frac{1}{\sqrt{2\pi}\sqrt{\beta^T\Sigma\beta}}\exp\left[-\frac{1}{2}\frac{(\beta^T x - \beta^T m)^2}{\beta^T\Sigma\beta}\right]$$

$$g = \frac{1}{\sqrt{2\pi}\sqrt{\beta^T\Sigma\beta}}\exp\left[-\frac{1}{2}\frac{(1 - \beta^T m)^2}{\beta^T\Sigma\beta}\right]$$

$$g_L = -\frac{1}{2}\frac{(1 - \beta^T m)}{\beta^T\Sigma\beta}$$

Also

$$Var\left(x \mid \beta^T x = 1\right) = E\left(xx^T \mid \beta^T x = 1\right) - \mu\mu^T$$

$$Var\left(x \mid \beta^T x = 1\right) = \Sigma - \frac{\Sigma\beta\beta^T\Sigma}{\beta^T\Sigma\beta}$$

So

$$V = E\left(xx^T \mid \beta^T x = 1\right) = \mu\mu^T + \Sigma - \frac{\Sigma\beta\beta^T\Sigma}{\beta^T\Sigma\beta}$$

Where μ is given from the expression above. These values can then be introduced in (3.26), (3.27), or (3.30) depending on the particular case.

Chapter 4

Cross-validation and Empirical Assessment of Estimates

4.1 Cross-validation corrections

When estimating parameters using data samples, there is always the risk of over-fitting. This is because the data is used both to estimate the parameter and to judge its performance, and therefore, gives a retrospective assessment of it. This means that the estimate of the parameter adapts to the particular data set used to obtain it. Therefore, when assessing its performance using the same data set, the results may be misleading, perhaps looking better than they would if using a different data set. Usually, this problem is overcome by carrying out a cross-validation process. This process helps eliminate the effect that the particular data set under study has on the estimation of the parameters.

The following idea for estimating cross-validation correction terms comes from Eguchi and Copas (2002). We start by judging the performance of the estimates obtained using the actual utility coming from the data set.

The retrospective assessment of $\hat{\beta}_h$, based on estimated utility, is given by

$$\hat{U} = \frac{1}{n}\sum_{i=1}^{n} I_{(0,\infty)}\left(\hat{\beta}_h^T x_i - 1\right)(y_i - c) \tag{4.1}$$

where the argument $\left(\hat{\beta}_h^T x_i - 1\right)$ of the indicator function decides for which observations the difference $(y_i - c)$ will be summed up. Now, suppose $\hat{\beta}_h$ is calculated from the same data, but removing the ith observation from the estimation process, giving $\hat{\beta}_h^{(i)}$. Then, the cross-validated version of (4.1) is

$$\hat{U}^{(CV)} = \frac{1}{n}\sum_{i=1}^{n} I_{(0,\infty)}\left(\hat{\beta}_h^{(i)T} x_i - 1\right)(y_i - c) \tag{4.2}$$

However, to calculate this in practice would represent a time and resource consuming process when the sample size is somewhat large. In this case, an approximation of the correction can be obtained following a procedure similar to the estimating process itself.

The idea is to start with the functions (3.10) and (3.13). In the latter, p' is substituted by its logistic regression estimate, so $p' = c(1-c)\left(\log\frac{c}{1-c} - \gamma_0\right) = bc(1-c)$ to give function

$$B_\beta = -\frac{bc(1-c)}{nh}\sum_{i=1}^{n} x_i x_i^T \left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{4.3}$$

Then the value of the functions removing the ith observation, denoted by $A_\beta^{(i)}$ and $B_\beta^{(i)}$

are

$$A_\beta^{(i)} = A_\beta - \frac{1}{nh} x_i (y_i - c)\, \phi\left(\frac{\beta^T x_i - 1}{h}\right) = A_\beta - \epsilon_i \tag{4.4}$$

$$B_\beta^{(i)} \simeq B_\beta - \left(-\frac{1}{nh} c(1-c)\, b x_i x_i^T \left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right)\right) \tag{4.5}$$

$$= B_\beta - D_i$$

Here ϵ_i and D_i give the part of the function corresponding to the ith observation, and which are removed from the sum. These values are of order n^{-1} and, therefore, relatively small. So, using the same argument as before for estimating β, the estimate for β removing the ith observation, $\beta_h^{(i)}$, is approximated with

$$\hat{\beta}_h^{(i)} \simeq \beta - (B_\beta - D_i)^{-1} (A_\beta - \epsilon_i)$$

And, since $\hat{\beta}_h = \beta - [B_\beta]^{-1} A_\beta$, then

$$\hat{\beta}_h^{(i)} - \hat{\beta}_h \simeq (B_\beta)^{-1} A_\beta - (B_\beta - D_i)^{-1} (A_\beta - \epsilon_i)$$

Let

$$(B_\beta - D_i)^{-1} = B_\beta^{-1} + E_i$$

From this expression, $(B_\beta - D_i)\left(B_\beta^{-1} + E_i\right) = I$. So considering E_i and D_i are small and consequently their product, it follows that $E_i \simeq B_\beta^{-1} D_i B_\beta^{-1}$. Using this approximation,

so that $(B_\beta - D_i)^{-1} \simeq B_\beta^{-1} + B_\beta^{-1} D_i B_\beta^{-1}$ and since $B_\beta^{-1} D_i B_\beta^{-1} \epsilon_i$ is also small, then

$$\begin{aligned}
\hat{\beta}_h^{(i)} - \hat{\beta}_h &\simeq (B_\beta)^{-1} A_\beta - \left(B_\beta{}^{-1} + B_\beta^{-1} D_i B_\beta^{-1}\right)(A_\beta - \epsilon_i) \\
\hat{\beta}_h^{(i)} - \hat{\beta}_h &\simeq B_\beta{}^{-1}\epsilon_i - B_\beta^{-1} D_i B_\beta^{-1} A_\beta \\
&= B_\beta{}^{-1}\left(\epsilon_i - D_i B_\beta^{-1} A_\beta\right)
\end{aligned}$$

Now, considering a difference between $\hat{\beta}_h^{(i)T} x_i$ and $\hat{\beta}_h^T x_i$ such that $\hat{\beta}_h^{(i)T} x_i = \hat{\beta}_h^T x_i + e_i$ and also that A_β is zero at the solution

$$e_i \simeq \epsilon_i B_\beta{}^{-1} x_i$$

and introducing the expression for ϵ_i then

$$e_i \simeq \frac{1}{nh}\phi\left(\frac{\beta^T x_i - 1}{h}\right)(y_i - c)\left(x_i^T B_\beta^{-1} x_i\right) \tag{4.6}$$

The cross-validated version of utility becomes.

$$\hat{U}^{(CV)} = \frac{1}{n}\sum_{i=1}^{n} I_{(0,\infty)}\left(\hat{\beta}_h^T x_i + e_i - 1\right)(y_i - c)$$

where e_i is given by (4.6). This will be used when comparing, for particular data sets, the non-parametric estimates and the estimates obtained using a parametric procedure.

4.1.1 Cross-validation correction for logistic regression estimates

Even though logistic regression estimates are asymptotically best and are, therefore, not really exposed to the dangers of over-fitting, here we will calculate a cross-validation correction term for them. Since comparisons between non-parametric and logistic regression estimates in different situations will be presented, this is just to standardise these comparisons.

In this case, the assumption is that the data arise from an underlying logistic model, that is $p_{x_i} = \Pr(y_i = 1|x_i)$ is given by

$$p_{x_i} = \frac{e^{\gamma_c^T x_{ci}}}{1 + e^{\gamma_c^T x_{ci}}}$$

where here γ_c denotes the vector of parameters including the intercept term γ_0 and $x_{ci} = (1, x_{1i}, x_{2i}, ..., x_{ri})$ is the vector of values of the r covariates for observation i including a 1 for the intercept term. The derivative of the log-likelihood function, which equalled to zero and solved for γ_c gives maximum likelihood estimators is

$$A_\gamma = \frac{\partial}{\partial \gamma_c} \log L_{\gamma_c} = \sum_{i=1}^{n} x_i (y_i - p_i)$$

and its derivative

$$A'_\gamma = B_\gamma = -\sum_{i=1}^{n} p_i (1 - p_i) x_i x_i^T$$

So, again, the versions of these functions where the ith observation is removed are

$$\begin{aligned} A_\gamma^{(i)} &= A_\gamma - x_i (y_i - p_i) = A_\gamma - \epsilon_i \\ B_\gamma^{(i)} &= B_\gamma - \left(-p_i (1 - p_i) x_i x_i^T\right) = B_\gamma - D_i \end{aligned}$$

where $\epsilon_i = x_i (y_i - p_i)$ and $D_i = -p_i (1 - p_i) x_i x_i^T$ are relatively small.

Now, γ_c and $\gamma_c^{(i)}$ are estimated in the following way

$$\begin{aligned} \hat{\gamma}_c &\simeq \gamma_c - B_\gamma^{-1} A_\gamma \\ \hat{\gamma}_c^{(i)} &\simeq \gamma_c - [B_\gamma - D_i]^{-1} [A_\gamma - \epsilon_i] \end{aligned}$$

So it follows that the difference between these two is

$$\hat{\gamma}_c^{(i)} - \hat{\gamma}_c \simeq B_\gamma^{-1} A_\gamma - [B_\gamma - D_i]^{-1} [A_\gamma - \epsilon_i] \tag{4.7}$$

To find an approximation for $(B_\gamma - D_i)^{-1}$ consider

$$(B_\gamma - D_i)^{-1} = B_\gamma^{-1} + E$$

$$(B_\gamma - D_i) \left(B_\gamma^{-1} + E\right) = I$$

Since $D_i E$ is small then

$$E \simeq B_\gamma^{-1} D_i B_\gamma^{-1}$$

And inserting this into (4.7) gives

$$\hat{\gamma}_c^{(i)} - \hat{\gamma}_c \simeq B_\gamma^{-1} A_\gamma - \left[B_\gamma^{-1} + B_\gamma^{-1} D_i B_\gamma^{-1}\right] \left[A_\gamma - \epsilon_i\right]$$

$$\hat{\gamma}_c^{(i)} \simeq \hat{\gamma}_c + B_\gamma^{-1} \epsilon_i$$

This approximation considers that $B_\gamma^{-1} D_i B_\gamma^{-1} \epsilon_i$ is small and that $B_\gamma^{-1} D_i B_\gamma^{-1} A_\gamma$ is zero because A_γ is zero at the solution for γ_c.

Also, in the notation used before

$$\hat{\beta}_{LR}^{(i)} = \frac{\hat{\gamma}^{(i)}}{\log\left(\frac{c}{1-c}\right) - \hat{\gamma}_0^{(i)}}$$

where here $\hat{\gamma}^{(i)} = (\gamma_1, \gamma_2, ..., \gamma_r)$ does not include the parameter estimate for the intercept term, denoted by $\hat{\gamma}_0^{(i)}$ in the same expression. This estimate for β is then introduced into the expression for calculating utility, giving

$$\hat{U}_{LR-mv}^{(CV)} \simeq \frac{1}{n} \sum_{i=1}^{n} I_{(0,\infty)} \left(\hat{\beta}_{LR}^{(i)T} x_i - 1\right) (y_i - c) \tag{4.8}$$

where $x_i = (x_{1i}, x_{2i}, ..., x_{ri})$

4.2 Behaviour of estimates and comparison with logistic regression estimates

Two examples to illustrate the theory presented above will be explored in this section. This will provide an idea of the behaviour of the estimates of the parameters and also how this is reflected in estimation of the utility, which is the ultimate objective. The examples are based on simulation of data and sample calculations of the quantities of interest. The first example explores these behaviours when the underlying model generating the data is logistic; the second example explores the same when the underlying model generating the data is an extreme value or Gumbel model. This will help illustrate the importance of using the appropriate model and the advantages of the non-parametric procedure when working with data which is clearly not logistic. As mentioned earlier there is an inevitable amount of over-fitting present at the moment of estimating, especially in the non-parametric procedure. Therefore, a comparison between non-crossvalidated and cross-validated estimates of utility is given.

Even though only the Gumbel model is presented for the case when the underlying process is not logistic, several other models were tested, both for the multivariate and the univariate case. These models included a biased logistic of the form

$$p_x = \left(\frac{e^{\gamma_0 + \gamma^T x}}{1 + e^{\gamma_0 + \gamma^T x}} \right)^{(1/\alpha)}$$

where different values of α were tried. A Weibull model of the form

$$p_x = \frac{1}{b}\left(\exp - \left(\frac{\gamma_0 + \gamma^T x - a}{b}\right)\right)$$

was also tried, giving different values to the parameters a and b. In every case, the non-parametric procedure had a better performance than logistic regression in terms of estimated utility. However, differences in this quantity were greater for models that were more biased with respect to the logistic one.

The sample size used for the examples presented is 20,000 observations. This is used because figures of this size are common in the credit application setting. It is important to mention that the results obtained are based on approximations that work well in the limit and therefore, a large sample size is assumed. However, smaller sample sizes were tried (although not reported). When the sample size is reduced, the variance automatically increases and therefore, the optimal window width h necessarily increases as well. This is translated in estimates which are less precise since the local procedure is expanded. The results for smaller sample sizes did not differ greatly from the results using 20,000 observations. However, when using very small sample sizes (for example, 500), the estimation procedure became somewhat unstable.

4.2.1 Example A - Logistic Generation

The settings for the first example are the following

$$
\begin{aligned}
x &\sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right] \\
\gamma_0 &= -1 \\
\gamma &= \begin{pmatrix} 1 \\ 0.5 \end{pmatrix} \\
p_x &= \frac{e^{\gamma_0+\gamma^T x}}{1+e^{\gamma_0+\gamma^T x}} = \frac{e^{-1+x_1+0.5x_2}}{1+e^{-1+x_1+0.5x_2}} \\
c &= 0.5 \\
\beta_{'true'} &= \begin{pmatrix} 1 \\ 0.5 \end{pmatrix} \\
\text{sample size} &= 20,000
\end{aligned}
$$

The optimal value for h, using this setting is around $h = 0.19$. According to these values, 1,000 samples of 20,000 observations were generated for each of 11 values of h, ranging from 0.15 to 0.35. Also, for each of these samples, 20,000 observations of a uniform random variable in the interval $[0, 1]$ were generated. Using p_x as above, the response variable y for each of the observations was assigned by making $y = 1$ for each observation where p_x was greater than the uniform random variable and $y = 0$ otherwise. Figure 4.2.1.1 presents the results of simulating 1,000 of these samples for each value of h and for each type of estimate (non-parametric and logistic regression). Plots (a) and (b) show non-parametric estimates of β_1 and β_2 are positively biased and that this

bias increases with h as was expected from the theory. They also show that the logistic regression estimates coincide almost exactly with the true value of the parameters, which was expected, since the process generating the data is logistic. Of course, the logistic regression estimates do not vary with h. The effects of overfitting are shown in Plot (c) of the same figure, where the non-parametric estimate of utility per 100,000 observations exceeds the logistic regression one for several values of h. Once the cross-validation correction is applied to the estimates, then utility for the non-parametric procedure decreases and goes below the one for logistic regression. In this example, some values of the non-parametric cross-validated estimate of utility go a little bit above the logistic regression estimate (which is also cross-validated). This is happening because different samples where taken at each value of h to calculate each estimate. However, in simulations where the same sample was used to calculate both estimates, the non-parametric cross-validated estimate of utility always fell below the logistic regression one. It is worth noting that while the cross-validation correction makes a big difference for the non-parametric estimate, it is hardly noticeable for the logistic regression one, which is in line with the theory.

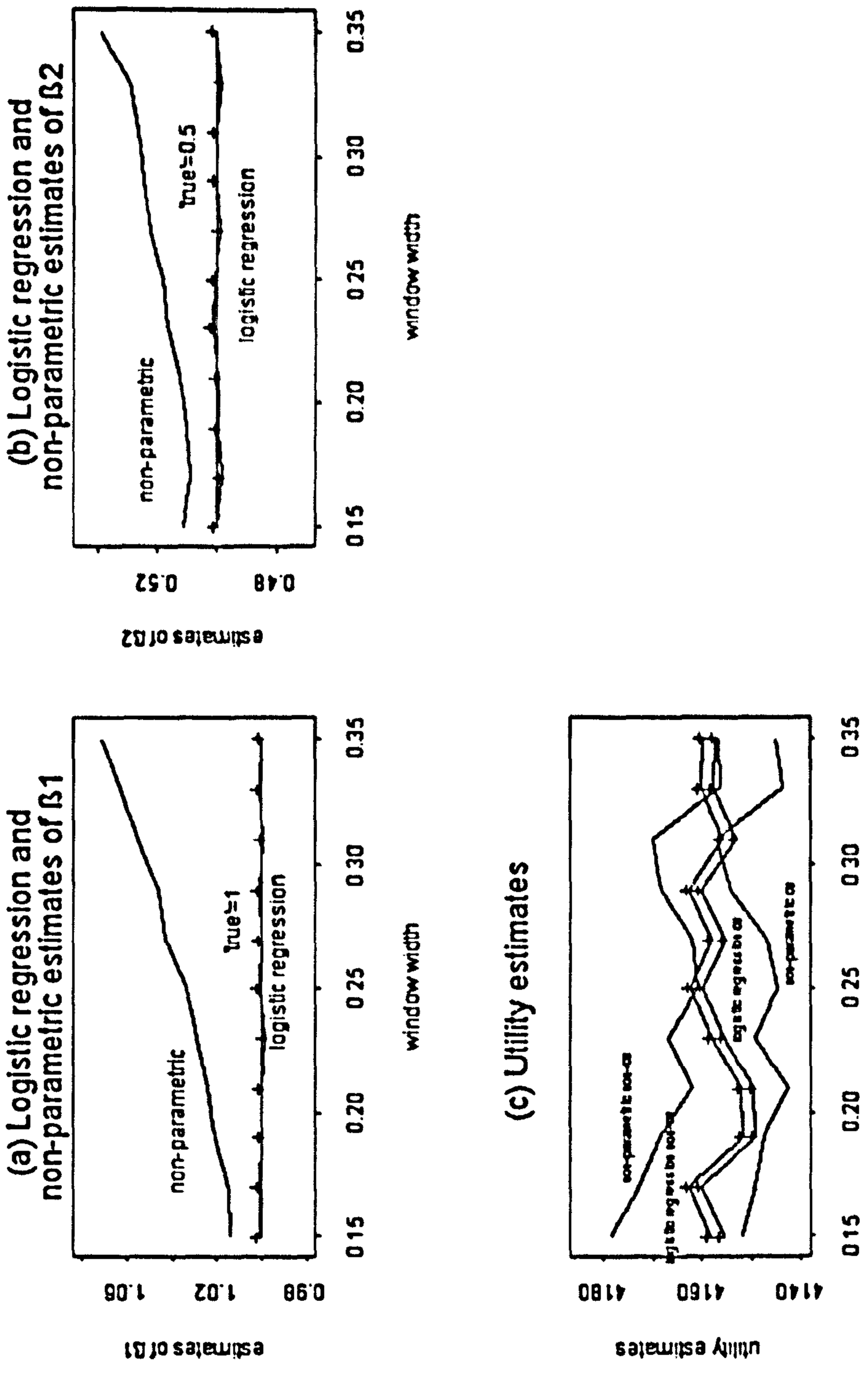

(a) Logistic regression and non-parametric estimates of β1
non-parametric
logistic regression
true=1
estimates of β1
window width
(b) Logistic regression and non-parametric estimates of β2
non-parametric
logistic regression
true=0.5
estimates of β2
window width
(c) Utility estimates
non-parametric
logistic regression
non-parametric
utility estimates
window width
utility per 100,000 observations

Figure 4.2.1.1

4.2.2 Example B - Gumbel Generation

The second example explores the behaviour of non-parametric and logistic regression estimates when the process generating the data is not logistic. Here a Gumbel model is assumed for p_x. The actual model used was

$$\begin{aligned} p_x &= \exp\left(-\exp\left(-\gamma_0 - \gamma^T x\right)\right) \\ \gamma_0 &= -1 \\ \gamma &= \begin{pmatrix} 1 \\ 0.5 \end{pmatrix} \end{aligned}$$

while everything else remained the same, including the size of the simulation process. However, the range of values of h used went from 0.18 to 0.33. Plots (a) and (b) of Figure 4.2.2.1 show the non-parametric and logistic regression estimates of β_1 and β_2 as well as the true values of the parameters ($\beta_1 = 0.7318$ and $\beta_2 = 0.3659$). In these cases, the non-parametric estimates are closer to the 'true' quantities for several values of h. Plot (c) shows cross-validated estimates of utility per 100,000 observations. It can be seen that the non-parametric estimate of utility exceeds the logistic regression estimate in every case, although this is more evident for values of h ranging between 0.225 and 0.285. It is worth mentioning that the calculation for optimal h gives a value of 0.266. This is obtained making use of (3.27) and from the plot, it looks as though the utility is actually larger around this point. Once again, it is evident that the cross-validation correction is larger for the non-parametric estimate of utility and also, logistic regression

estimates do not vary over the range of h, except due to sampling effects.

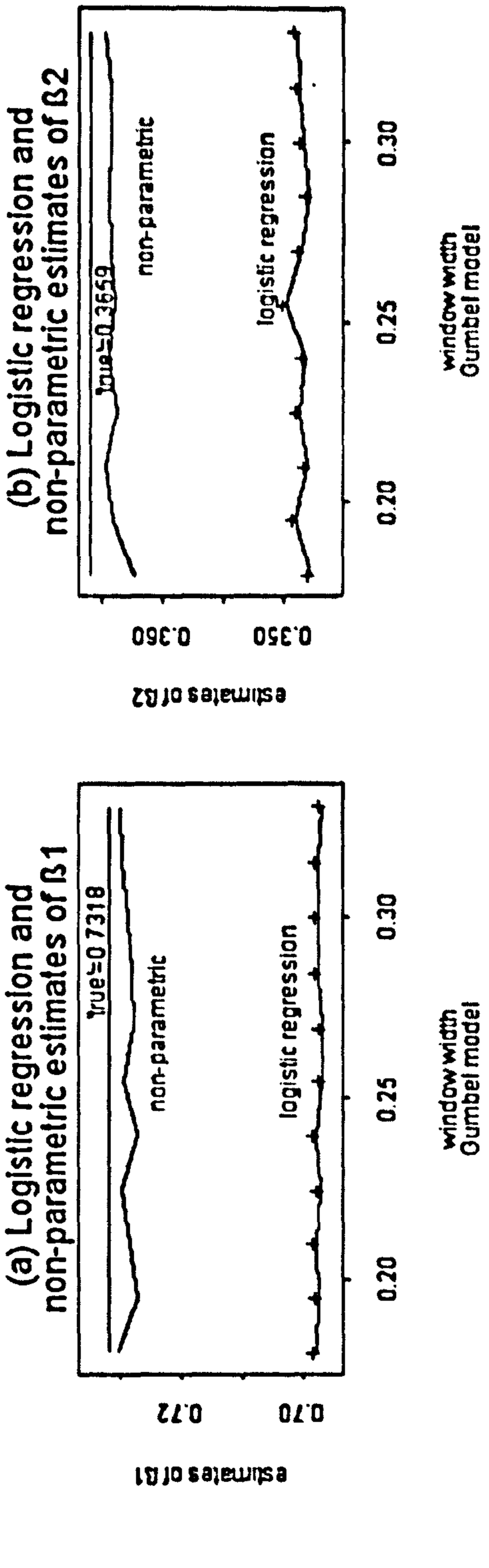


Figure 4.2.2.1

Chapter 5

Univariate Case

5.1 Parametrisation and Estimation

This chapter refers to the case of only one covariate. The parametrisation used here was somewhat different. This is because with only one covariate, there is only need for one parameter which coincides with the cut-off point. This limit, say k, is a point above or below which all individuals will be considered (accepted, treated, etc.) and the coefficient on x becomes redundant. This parametrisation, however, is completely analogous to the one described in §3 for the case of only one covariate. In that case the utility function is given by

$$U_\beta = \int_{\beta x > 1} (p_x - c)\, f_x dx$$

implying that the integral is calculated on the interval $x > 1/\beta$ ($\beta \neq 0$). Using the previous results for calculating the derivative for U_β, this would be

$$\frac{\partial}{\partial\beta}U_\beta = \frac{1}{\beta^2}\left(p_{1/\beta} - c\right) f_{1/\beta}$$

and maximum utility would be obtained at the root in $1/\beta$ of

$$p_{1/\beta} = c.$$

For example, in the logistic, setting this would mean that maximum utility is obtained at

$$\frac{1}{\beta} = \frac{\log\left(\frac{c}{1-c}\right) - \gamma_0}{\gamma_1} \text{ or}$$

$$\beta = \frac{\gamma_1}{\log\left(\frac{c}{1-c}\right) - \gamma_0}$$

where $p_x = \frac{\exp(\gamma_0+\gamma_1 x)}{1+\exp(\gamma_0+\gamma_1 x)}$. This is equivalent to the result obtained in the present chapter, as will be seen, where $k = 1/\beta$. All the calculations are repeated because the case of one covariate is much simpler and has a direct application in quantal bioassay, so it is worth discussing. Also, the implications of §7 are obviously not present here, since the estimate of k is optimal for utility and, therefore, it includes any optimal relocation of x.

The utility function is now

$$U_k = \int_{x>k} (p_x - c)\, f_x dx \tag{5.1}$$

Parametrising in this way has certain advantages that will become evident as we go along. The first is that the derivatives and the maximising procedure are much simpler.

Here, derivatives are taken with respect to k.

$$\frac{\partial}{\partial k} U_k = -(p_k - c)\, f_k \tag{5.2}$$

So, in this case, the maximum utility is found when k is such that $p_k = c$. For example, if p_x is given by the logistic model with parameters γ_0 and γ_1 then

$$k_{opt} = \frac{\log\left(\frac{c}{1-c}\right) - \gamma_0}{\gamma_1}$$

Since usually p_x is not known, a procedure similar to the one mentioned in the multivariate case will be used to estimate k. Let us define A_k as

$$A_k = \frac{1}{nh} \sum_{i=1}^{n} (y_i - c)\, \phi\left(\frac{x_i - k}{h}\right) \tag{5.3}$$

where the kernel is defined by $\phi\left(\frac{x_i - k}{h}\right)$, the standard normal density function evaluated

at $\frac{x_i-k}{h}$. The expectation of A_k is

$$E_y(A_k) = \frac{1}{nh}\sum_{i=1}^{n}(p_x - c)\,\phi\left(\frac{x_i - k}{h}\right) \tag{5.4}$$

In the sum, the largest weights, $\phi\left(\frac{x_i-k}{h}\right)$, will be given to the observations which are close to k and (5.4) will become zero at the value of k where p_x is equal to c. In this sense, A_k is just a non-parametric approximation to (5.2). If k is a root of A_k then $\hat{k}_h \simeq k - [E_y(A'_k)]^{-1} A_k$, and $\hat{k}_h$ will give a non-parametric estimate of k. Again, the expectation of A'_k is taken to make the estimation process more stable. Also, the subindex h in $\hat{k}_h$ indicates that the estimate of k depends on the value of the window width (h). Here A'_k, and $E_y(A'_k)$ are given by

$$A'_k = \frac{1}{nh^2}\sum_{i=1}^{n}(y_i - c)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right) \tag{5.5}$$

$$E_y(A'_k) = \frac{1}{nh^2}\sum_{i=1}^{n}(p_x - c)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right)$$

And expanding p_x around k, keeping only first order terms will give

$$E_y(A'_k) \simeq \frac{1}{nh}p'_k\sum_{i=1}^{n}\left(\frac{x_i - k}{h}\right)^2\phi\left(\frac{x_i - k}{h}\right) \tag{5.6}$$

where p'_k denotes the first derivative of $p_x$with respect to x, evaluated at the value k. So (5.3) and (5.6) are the functions that will be used in the non-parametric estimation procedure. In such procedure, the logistic regression estimate of p'_k is used.

5.2 Approximations for bias and variance of $\hat{k}_h$

It is important to calculate the bias and variance of $\hat{k}_h$ because the final objective is to choose a $\hat{k}_h$ in order to maximise $E\left(U_{\hat{k}_h}\right)$ and this will depend on the choice of h. The bias and the variance are obtained so that an optimal value of the window width h can be approximated and values around this figure tested in order to search for the estimate of k that gives the maximum utility.

We have that

$$Bias\left(\hat{k}_h\right) = E_y\left[\hat{k}_h - k\right] \simeq -\left[E_y\left(A'_k\right)\right]^{-1} E_y\left(A_k\right) \tag{5.7}$$

$$Var\left(\hat{k}_h\right) \simeq \left[E_y\left(A'_k\right)\right]^{-2} Var_y\left(A_k\right) \tag{5.8}$$

To find approximations for $E_y\left(A_k\right)$, $E_y\left(A'_k\right)$ and $Var_y\left(A_k\right)$ Taylor expansions will be used. Starting with $E_y\left(A_k\right)$, we have

$$\begin{aligned} E_y\left(A_k\right) &= \frac{1}{nh}\sum_{i=1}^{n}\left(p_x - c\right)\phi\left(\frac{x_i - k}{h}\right) \\ &\simeq \frac{1}{h}\int_x \left(p_x - c\right)\phi\left(\frac{x - k}{h}\right) f_x dx \end{aligned}$$

If we do the variable change $u = \frac{x-k}{h}$ and expand, keeping only first order terms, this

will become

$$\begin{aligned} E_y(A_k) &\simeq \int_u \left(p + hup' + \frac{1}{2}h^2u^2p'' - c\right)(f + huf')\,\phi(u)\,du \\ &\simeq h^2\left[p'f' + \frac{1}{2}p''f\right] \end{aligned}$$

In the expressions above, the lack of suffix in the functions p and f, and their derivatives indicates that they are evaluated at k. Also, at k, p is equal to c. For $E_y(A'_k)$ we have

$$\begin{aligned} E_y(A'_k) &= \frac{1}{nh^2}\sum_{i=1}^{n}(p_x - c)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right) \\ &\simeq \frac{1}{h^2}\int_x (p_x - c)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x - k}{h}\right) f_x dx \end{aligned}$$

Again changing variable and expanding gives

$$E_y(A'_k) \simeq \frac{1}{h}\int_u (p + hup' - c)(f + huf')\,u\phi_u du \simeq p'f$$

So the bias of $\hat{k}_h$ is given by

$$Bias\left(\hat{k}_h\right) \simeq -h^2\left(\frac{1}{p'f}\right)\left(p'f' + \frac{1}{2}p''f\right) \simeq -h^2\left(f_L + \frac{1}{2}\frac{p''}{p'}\right) \tag{5.9}$$

where

$$\begin{aligned} f_L &= \frac{d}{dx}\log f_x \mid_{x=k} \\ p' &= \frac{d}{dx}p_x \mid_{x=k} \\ p'' &= \frac{d^2}{dx^2}p_x \mid_{x=k} \end{aligned}$$

The variance of A_k is given by

$$\begin{aligned} Var_y(A_k) &= \frac{1}{n^2h^2}\sum_{i=1}^{n} p_x(1-p_x)\phi^2\left(\frac{x_i-k}{h}\right) \\ &\simeq \frac{1}{nh^2}\int_x p_x(1-p_x)\phi^2\left(\frac{x-k}{h}\right)f_x dx \end{aligned}$$

Now the variable change is: $u = \sqrt{2}\left(\frac{x-k}{h}\right)$ and, after expanding, the variance is approximated as

$$\begin{aligned} Var_y(A_k) &\simeq \frac{1}{nh}\frac{1}{\sqrt{4\pi}}\int_u p(t_u)(1-p(t_u))f(t_u)\phi_u du \\ &\simeq \frac{1}{nh}\frac{1}{\sqrt{4\pi}}\int_u c(1-c)f\phi_u du \\ &\simeq \frac{1}{nh}\frac{1}{\sqrt{4\pi}}c(1-c)f \end{aligned}$$

where $t_u = \frac{1}{\sqrt{2}}hu + k$. So finally, the variance of $\hat{k}_h$ is given by

$$Var\left(\hat{k}_h\right) \simeq \frac{1}{(p'f)^2}\frac{1}{nh}\frac{1}{\sqrt{4\pi}}c(1-c)f = \frac{1}{nh}\frac{1}{\sqrt{4\pi}}\frac{c(1-c)}{(p')^2 f} \tag{5.10}$$

5.3 Optimal window width h

The main objective is to maximise expected utility, so this is the idea that will be used to calculate the optimal window width. We already have the first derivative of utility (5.2), the second derivative is given by

$$\frac{\partial^2}{\partial k^2}(U_k) \simeq -p'f \tag{5.11}$$

Utility evaluated at $\hat{k}_h$ may be expanded around k as

$$U_{\hat{k}} \simeq U_k + \left(\hat{k} - k\right) U_k' + \frac{1}{2}\left(\hat{k} - k\right)^2 U_k''$$

Since U_k' is zero, the expectation is approximated as

$$E\left(U_{\hat{k}}\right) \simeq U_k + \frac{1}{2} E\left[\left(\hat{k} - k\right)^2\right] U_k''$$

And since

$$\begin{aligned} Var\left(\hat{k}_h\right) &\simeq E\left(\hat{k}_h^2\right) - E^2\left(\hat{k}_h\right) \\ Bias\left(\hat{k}_h\right) &= E\left(\hat{k}_h - k\right) = E\left(\hat{k}_h\right) - k \end{aligned}$$

Then, expected utility is approximated as

$$E(U_k) \simeq U_k + \frac{1}{2}U_k'' \left[Var\left(\hat{k}_h\right) + Bias^2\left(\hat{k}_h\right) \right]$$

Let

$$\begin{aligned} Bias\left(\hat{k}_h\right) &\simeq -h^2\alpha_u \\ Var\left(\hat{k}_h\right) &\simeq \frac{1}{nh}B_u \end{aligned}$$

where

$$\begin{aligned} \alpha_u &= f_L + \frac{1}{2}\frac{p''}{p'} \\ B_u &= \frac{c(1-c)}{\sqrt{4\pi}\,(p')^2 f} \end{aligned}$$

To maximise expected utility, we take the derivative with respect to h and equal to zero. This gives

$$\begin{aligned} \frac{\partial}{\partial h}E(U_k) &\simeq 2\alpha_u^2 h^3 U_k'' - \frac{1}{2nh^2}B_u U_k'' \\ 0 &= 4\alpha_u^2 h^5 U_k'' - \frac{1}{n}B_u U_k'' \end{aligned}$$

And solving for h gives

$$h_{opt\_univ} \simeq \left\{ \frac{1}{4n}\frac{B_u}{\alpha_u^2} \right\}^{1/5} \propto n^{-1/5} \tag{5.12}$$

Here, the optimal h is again of same order of magnitude as the optimal window width found by Silverman (1986) for density estimation using kernels. These expressions can also be simplified for the special cases when the probability of success is a function of the score, when probability of success is given by a logistic function, and when the covariate is normally distributed.

5.4 Link to quantal bioassay

Quantal bioassays are a type of study carried out in fields such as toxicology and pharmacology. The aim is to assess the toxic effect of substances or the effective action of a drug or vaccine. In these types of study, deaths or other binary responses are measured after the subjects have been exposed to the drug in study at different dose levels. Usually, the median effective dose is sought (ED50). This is the dose level at which 50% of the subjects have reached the status looked for (death, development of tumour, etc.).

In quantal bioassays, the underlying assumption is that the observed reaction y_i of the ith subject (for example $y_i = 0$ no reaction, $y_i = 1$ reaction) at the log dose level x_i results from a Bernoulli trial with probability p_{x_i} and that these Bernoulli trials are independent for different subjects. Here, $p : R \rightarrow [0, 1]$ denotes the dose-response curve.

The distribution for y_i is given by

$$\begin{aligned} \Pr(y_i = 1|x_i) &= p_{x_i} \\ \Pr(y_i = 0|x_i) &= 1 - p_{x_i} \\ i &= 1, \ldots, n \end{aligned}$$

The log effective doses at which $100\alpha\%$ of individuals react are called the functionals of p. Another assumption is that p_x is strictly monotone increasing (approaching 0 and 1 as asymptotes) in order that its specific functionals are well defined for $0 < \alpha < 1$. The estimation of the curve p and of the functionals of p is the aim of the statistical analysis in quantal bioassay. The usual approach for the estimation of the functionals is

$$\left(\log \hat{E}D\alpha\right) = \hat{p}^{-1}(\alpha)$$

where p^{-1} is the inverse function of the dose-response curve.

So once the dose-response curve p is obtained, the functionals are estimated from it. Both parametric and non-parametric approaches have been considered in order to estimate p. The most commonly used parametric approaches involve the use of logit and probit models. In the parametric setting, the method of maximum likelihood generally used for estimation reduces the infinite dimensional problem to only a few parameters, and yields asymptotic confidence intervals. However, the efficiency of the method is centred on the assumption that the true curve follows the assumed model.

There are mechanisms such as the biological effects of a drug action or credit risk behaviour which may follow patterns that can be somewhat biased when compared to logit or probit models. For cases where the dose-response curve is mostly unknown, non-parametric methods can provide a good alternative.

A number of semi-parametric and non-parametric procedures have been used to model quantal bioassay information, see for example Taylor (1995). or Tsodikov et al (1995). In particular, generalized additive models as proposed by Hastie and Tibshirani (1990), are widely mentioned in the literature. The locally-weighted smoother of Cleveland (1979) which is currently called *loess* in the S statistical computing language can be applied to any polynomial, but local lines are more generally used. Locally-weighted smoothers are popular since they enjoy the best of two worlds. On one hand, they share the ability of near-neighbour smoothers to adapt their bandwidth to the local density of the predictors, and on the other, they have the smoothness features of kernel smoothers. In a paper by Fan et al (1995), local polynomial kernel regression is applied to a data set consisting of dispositions of burn victims and several covariates. In this paper, a local linear estimate is used to estimate the probability of survival from burns using a transformation of the area of third-degree burn. This application can be related to quantal assay.

Regarding the parametric procedures used to tackle this problem, probit and logit models have been used widely, as mentioned earlier. However, a slightly different approach may be found in Pack et al (1990).

The problem of estimating functionals in quantal bioassay is essentially the same as

finding the root of (5.3) as described above. In this case, the value of c stands for the value of α. The estimation is concentrated around this value and all the observations are weighed in order to be used in the calculation. Since the estimation of only one point, and not of the complete dose-response curve p is of interest, the results should be, in principle, more precise.

In a paper by Müller and Schmitt (1988) a non-parametric kernel estimator for the dose-response curve is proposed

$$\hat{p}_x = \frac{1}{b}\sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{b}\right) du \tag{5.13}$$

where $s_0 = 0$, $s_n = 1$, $s_i = \frac{1}{2}(d_i + d_{i+1})$, $1 \leq i \leq n-1$ is an interpolating sequence of the doses denoted by d_i, b is the window width, and K is any continuous kernel function satisfying

$$\int K(v)\,dv = 1, \int K(v)\,v dv = 0, \int K(v)\,v^2 dv > 0$$

Also, support$(K) = [-1, 1]$, $b \to 0$, and $nb \to \infty$ as $n \to \infty$.

We found that the behaviour of this estimator when the doses are not equidistant is not optimal, and this is only because of the type of kernel function used, as it is the only difference between both approaches. This can be seen doing the following transformation. If in (5.13) the quantity c is subtracted from both sides of the equation, a function equivalent to A_k as in (5.3) is found. Recalling (5.3) and taking $\hat{p}_x - c$ in

(5.13) above gives

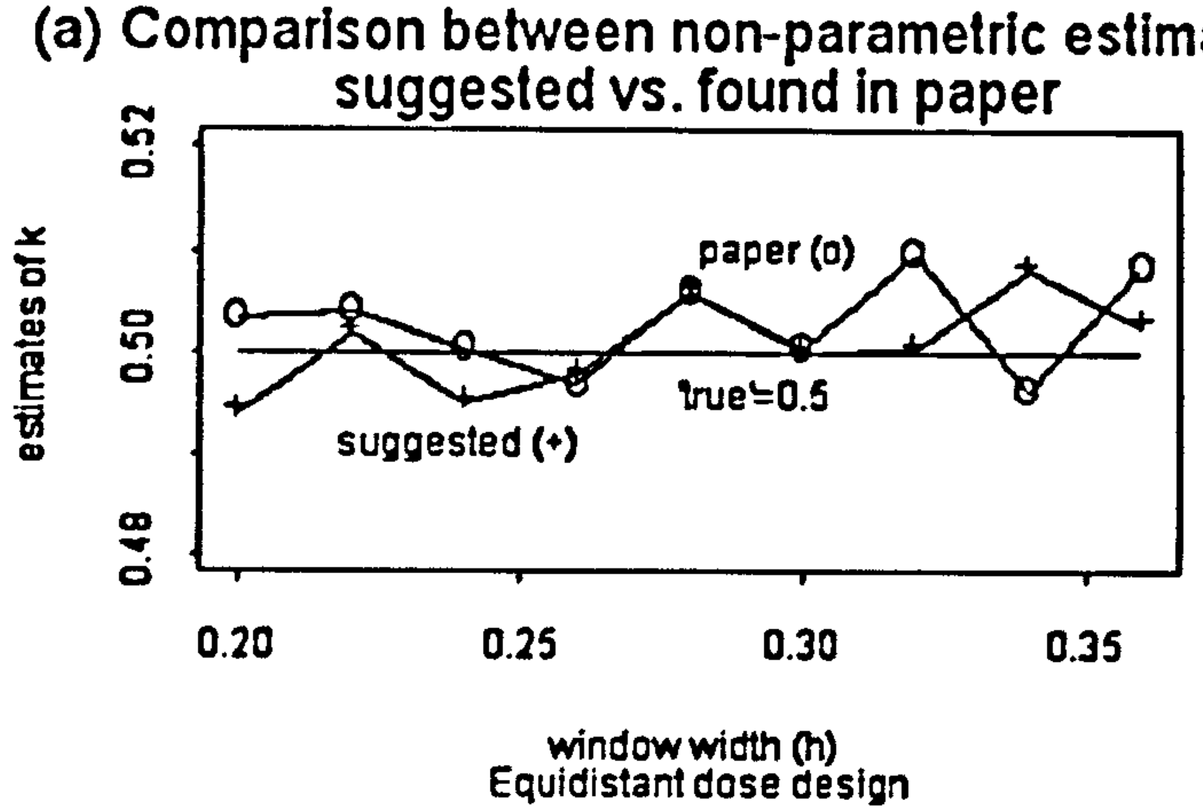
$$
\begin{aligned}
A_k &= \frac{1}{nh}\sum_{i=1}^{n}(y_i - c)\,\phi\left(\frac{x_i - k}{h}\right) \\
A_k^* &= \hat{p}_x - c = \frac{1}{\Phi\left(\frac{1-k}{h}\right) - \Phi\left(-\frac{k}{h}\right)} \times \\
&\qquad \sum_{i=1}^{n}(y_i - c)\left[\Phi\left(\frac{s_i - k}{h}\right) - \Phi\left(\frac{s_{i-1} - k}{h}\right)\right]
\end{aligned}
$$

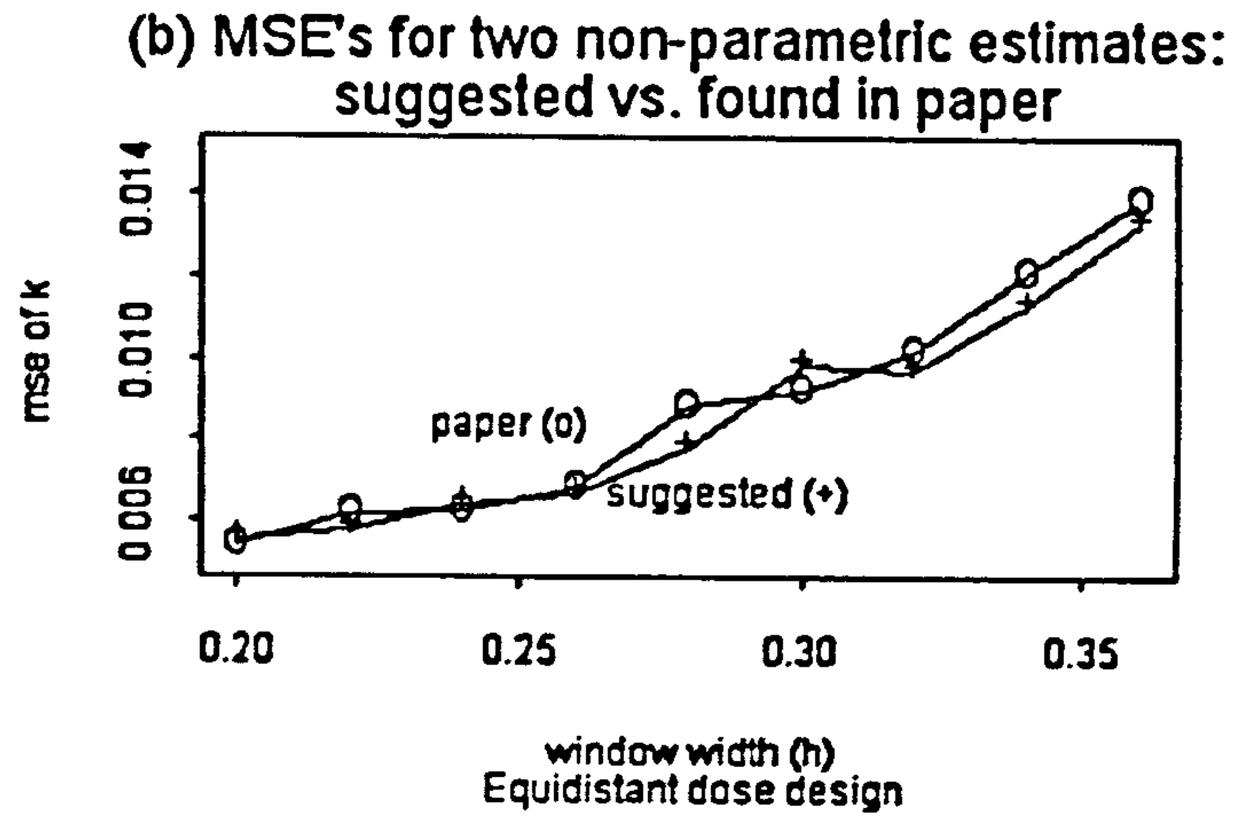
where $s_0 = 0$, $s_n = 1$, $s_i = \frac{1}{2}(x_i + x_{i+1})$, $1 \leq i \leq n-1$ is again an interpolating sequence of the doses, and $\Phi(u)$ is the standard normal cumulative distribution function. The variable d (for the doses) has been substituted by x. These two functions are now equivalent and the value of k that makes each of them zero will be the quantity we are looking for. The only difference between them is the type of kernel function used, since A_k^* uses the interpolating sequence of doses. This is why the cumulative distribution function is used instead of the probability density function. Both functions are normalised so that the sum of all the kernel weights adds up to one. However, this is not necessary because the value of k that makes the functions zero will be the same independently of the scaling factor. Even though the mentioned paper uses kernel functions with support$(K) = [-1, 1]$ a slight adaptation in this sense is made to make the results more comparable and this is why gaussian kernels are used.

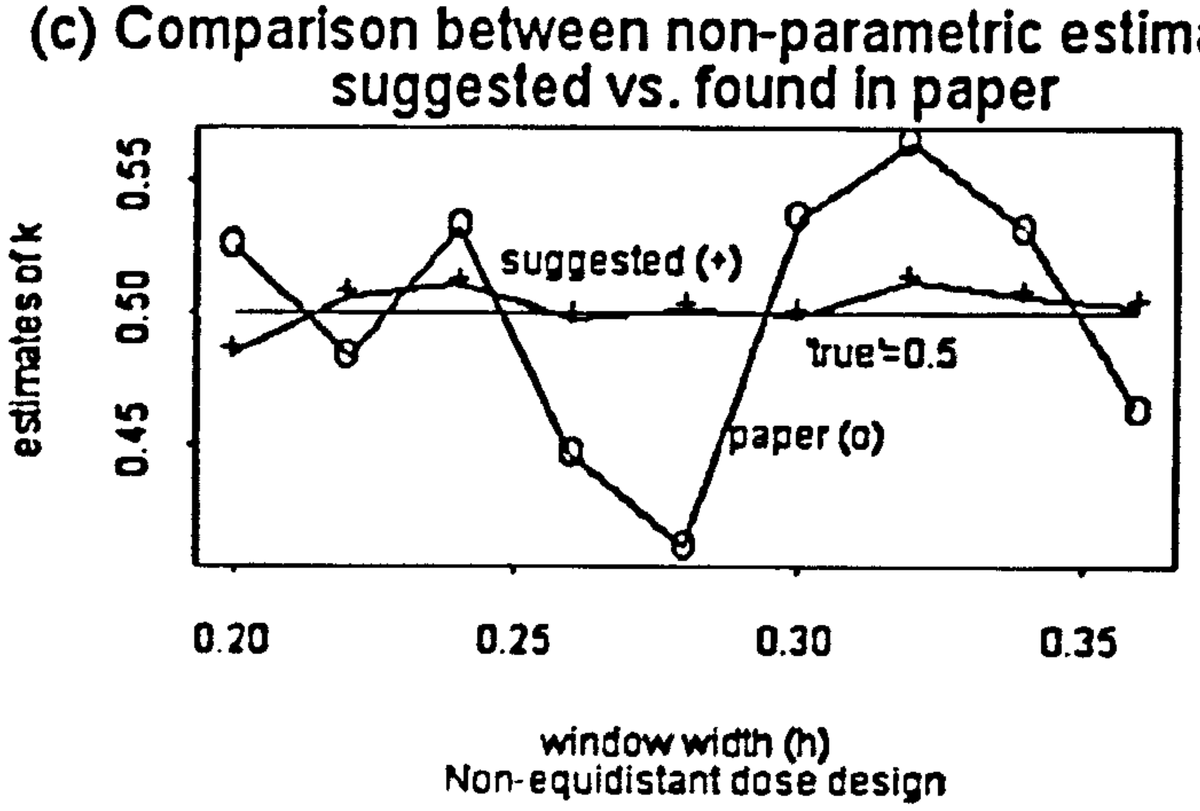
To compare the performance of these functions, two artificial data sets were produced. A logistic model for p_x with parameters $\gamma_0 = -2$ and $\gamma_1 = 4$ determined the probability of success in 100 observations of log doses (called x). Two different designs were used,

the first one considered equally spaced doses where $x = \frac{i-1}{n-1}$ for $i = 1, .., 100$ and $n = 100$. The second one considered a non-equidistant spacing, making a quarter of the set of log doses smaller than 0.025, a quarter of this set greater than 0.975 and half of the set ranging from 0.475 to 0.525. For this design, another 100 observations were considered. The response was obtained by assigning a value of 1 to the doses with p_x greater than a uniform random variable in the interval $[0, 1]$ and a value of 0 otherwise. A range of values of h $(0.20, 0.225, ..., 0.35, 0.375)$ was used in order to observe differences in the estimates depending on this value. Under this logistic model, the smallest probability of success is around 0.12 while the largest is around 0.88.

The first thing to note is that if c is 0.5, then the value of k that solves both A_k and A_k^* is around 0.5. Figure 5.4.1 shows that for the equidistant spacing design, there is no real difference in the performance of the estimates of k that each function produces. However, in the non-equidistant spacing design the estimate of k obtained using A_k^* for the estimation has a larger variance with respect to the estimate given by A_k. Consequently, this is also reflected in the mean squared error of the estimate, which is considerably larger than the mean squared error for $\hat{k}_h$ using A_k. This exercise was performed in order to point out that a kernel function of the type used in A_k is more adequate, independently of the distribution of the covariate under study.


(a) Comparison between non-parametric estimate:
suggested vs. found in paper
estimates of k
0.48
0.50
0.52
paper (o)
true=0.5
suggested (+)
0.20
0.25
0.30
0.35
window width (h)
Equidistant dose design


(b) MSE's for two non-parametric estimates:
suggested vs. found in paper
mse of k
0.006
0.010
0.014
paper (o)
suggested (+)
0.20
0.25
0.30
0.35
window width (h)
Equidistant dose design


(c) Comparison between non-parametric estimate:
suggested vs. found in paper
estimates of k
0.45
0.50
0.55
suggested (+)
true=0.5
paper (o)
0.20
0.25
0.30
0.35
window width (h)
Non-equidistant dose design

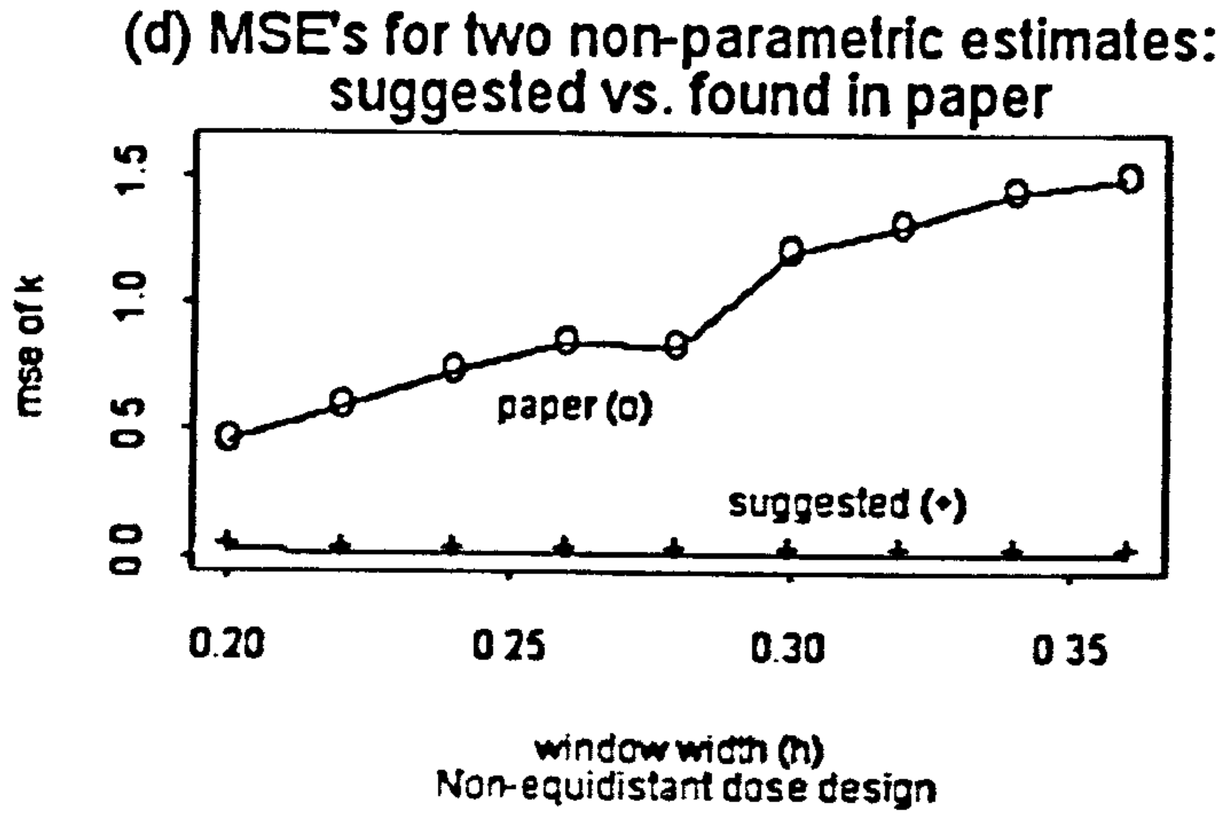

(d) MSE's for two non-parametric estimates:
suggested vs. found in paper
mse of k
0.0
0.5
1.0
1.5
paper (o)
suggested (+)
0.20
0.25
0.30
0.35
window width (h)
Non-equidistant dose design

Figure 5.4.1

5.5 Cross-validation corrections

This section explores the calculation of cross-validation corrections for the estimate of k using the same idea as for the case of multivariate x seen before. Again, we start by judging the performance of the estimates from the actual utility obtained from the data set.

The retrospective assessment of k_h based on estimated utility is given by

$$\hat{U}_u = \sum_{i=1}^{n} I_{(0,\infty)} \left(x_i - \hat{k} \right) (y_i - c) \tag{5.14}$$

where the subindex in $\hat{U}_u$ indicates (5.14) is a univariate estimate of utility. Here the argument $\left(x_i - \hat{k} \right)$ of the indicator function decides for which observations the difference $(y_i - c)$ will be summed up. Now, suppose $\hat{k}$ is calculated from the same data, but removing the ith observation from the estimation process, giving $\hat{k}^{(i)}$. Then, the cross-validated version of (5.14) is

$$\hat{U}_u^{(CV)} = \sum_{i=1}^{n} I_{(0,\infty)} \left(x_i - \hat{k}^{(i)} \right) (y_i - c) \tag{5.15}$$

Since this process is difficult in practice when the sample size is somewhat large, an approximation of the correction can be obtained, following a procedure similar to the estimating process itself. The idea, as before, is to start with the functions (5.3) and (5.6). In the latter, p'_k is substituted by its logistic regression estimate, so $p'_k = c\,(1-c)\,\gamma$ to give function B_k. From these, the value of the functions removing the ith observation,

denoted by $A_k^{(i)}$ and $B_k^{(i)}$ are

$$A_k^{(i)} = A_k - \frac{1}{nh}(y_i - c)\,\phi\left(\frac{x_i - k}{h}\right) = A_k - \epsilon_i \tag{5.16}$$

$$B_k^{(i)} \simeq B_k - \frac{1}{nh}c(1-c)\,\gamma\left(\frac{x_i - k}{h}\right)^2 \phi\left(\frac{x_i - k}{h}\right) = B_k - D_i \tag{5.17}$$

where ϵ_i and D_i give the part of the function corresponding to the ith observation, and which are removed from the sum. So, using the same argument as before for estimating k, the estimate for k removing the ith observation, $\hat{k}_h^{(i)}$ is

$$\hat{k}_h^{(i)} \simeq k - \frac{A_k - \epsilon_i}{B_k - D_i}$$

And, since $\hat{k}_h \simeq k - \frac{A_k}{B_k}$, then

$$\hat{k}_h^{(i)} \simeq \hat{k}_h + \frac{A_k}{B_k} - \frac{A_k - \epsilon_i}{B_k - D_i}$$

Let

$$\frac{1}{B_k - D_i} = \frac{1}{B_k} + E_i$$

From this expression, considering E_i and D_i are small and therefore their product, it follows that $E_i \simeq \frac{D_i}{B_k^2}$ and

$$\hat{k}_h^{(i)} \simeq \hat{k}_h + \frac{A_k}{B_k} - \left(\frac{1}{B_k} + \frac{D_i}{B_k^2}\right)(A_k - \epsilon_i)$$

Since $\frac{\epsilon_i D_i}{B_k^2}$ is again small and A_k is zero at the solution by definition, then

$$\hat{k}_h^{(i)} \simeq \hat{k}_h + \frac{1}{B_k}\epsilon_i$$

So finally

$$\hat{k}_h^{(i)} \simeq \hat{k}_h + B_k^{-1}\left[\frac{1}{nh}(y_i - c)\,\phi\left(\frac{x_i - k}{h}\right)\right]$$

When $\hat{k}_h^{(i)}$ as above is used in (5.15) then the cross-validated approximation to utility is obtained. This will be used when comparing, for particular data sets, the non-parametric estimates and the estimates obtained using a parametric procedure.

To obtain the analogue of a cross-validation correction for the logistic regression estimate in the univariate parametrisation, again the assumption is that the data arise from an underlying logistic model, that is $p_{x_i} = \Pr(y_i = 1|x_i)$ is given by

$$p_{x_i} = \frac{e^{\gamma_0+\gamma_1 x_i}}{1 + e^{\gamma_0+\gamma_1 x_i}}$$

The derivative of the log-likelihood function, which equalled to zero and solved for γ_0 and γ_1 gives maximum likelihood estimators is

$$A_\gamma = \frac{\partial}{\partial \gamma_c}\log L_\gamma = \sum x_i\,(y_i - p_i)$$

where here $\gamma_c^T = (\gamma_0, \gamma_1)$ and $x_i = (1, x_i)$ for each observation. The derivative of A_γ is

given by

$$A'_\gamma = B_\gamma = -\sum p_i (1 - p_i) x_i x_i^T$$

These functions are just the same as the ones used in the multivariate case and to obtain the estimate for $\hat{\gamma}_c^{(i)}$ exactly the same steps are followed. So, again,

$$\hat{\gamma}_c^{(i)} \simeq \hat{\gamma}_c + B_\gamma^{-1} \epsilon_i$$

where here $\epsilon_i = x_i (y_i - p_i)$. Also, in the notation used before

$$\hat{k}_{LR}^{(i)} = \frac{\log\left(\frac{c}{1-c}\right) - \hat{\gamma}_0^{(i)}}{\hat{\gamma}^{(i)}}$$

where here $\hat{\gamma}_1^{(i)}$ does not include the intercept term, denoted by $\hat{\gamma}_0^{(i)}$ in the same expression.

This estimate for k is then introduced in the expression for calculating utility, giving

$$\hat{U}_{LR-u}^{(CV)} \simeq \frac{1}{n} \sum_{i=1}^{n} I_{(0,\infty)} \left(x_i - \hat{k}_{LR}^{(i)} \right) (y_i - c) \tag{5.18}$$

5.6 Behaviour of estimates and comparison with logistic regression estimates under different underlying models

In this section, some examples to illustrate the theory presented above will be explored. This will help in two ways, the first one will be to obtain an empirical assessment of the results and it will also provide a comparison against traditional procedures. As a complement, an evaluation on the goodness of the non-parametric method will be obtained. The idea will be to artificially produce an example and then observe what happens when the results presented above are applied to it. A simulation procedure will be carried out and estimates of variance, bias and mean squared error of $\hat{k}_h$ will be compared to the theory presented here.

The simulation procedure will also help compare the behaviour of logistic regression estimates and estimates obtained by the non-parametric procedure. As mentioned earlier, some over-fitting may occur when estimating. Therefore, a comparison between non-crossvalidated and cross-validated estimates is also given. The simulation process will consider two different situations. For the first one, a logistic underlying model will be used, and for the second one a Gumbel, and consequently biased, model will generate the data. This will help illustrate the importance of using the appropriate model. When the process generating the data is relatively close to a logistic procedure, then this is always the best model to use. However, when the data do not correspond more or less to

a logistic generating process, using this model may produce sub-optimal estimates. For these cases a non-parametric procedure such as the one suggested here would represent a better choice.

5.6.1 Example A - Logistic Generation

The settings for the first example are the following

$$\begin{aligned} x &\sim N(0,1) \\ \gamma_0 &= -0.5 \text{ and } \gamma_1 = 0.5 \\ p_x &= \frac{e^{\gamma_0+\gamma_1 x}}{1+e^{\gamma_0+\gamma_1 x}} = \frac{e^{-0.5+0.5x}}{1+e^{-0.5+0.5x}} \\ c &= 0.5 \\ k_{'true'} &= 1 \\ \text{sample size} &= 20,000 \end{aligned}$$

For Plot (a) of Figure 5.6.1.1, 200 samples of 20,000 observations having $x \sim N(0,1)$ are obtained for each of 17 values of k, the cutoff

$$k = (-0.2, 0, 0.2, ..., 2.6, 2.8, 3.0)$$

Responses are simulated by generating a uniform random variable in the interval $[0,1]$ and assigning $y = 1$ if p_{x_i} (applying the logistic model given to each individual in the

sample) is greater than this variable and $y = 0$ otherwise. The average utility, calculated using (5.14), of these 200 samples is plotted against the values of k. It can be seen that this is a smooth function, achieving a unique maximum and that this maximum is around $1,020$ units per $100,000$ observations at $k = 1$. For the credit card example this means that for every $100,000$ applications where only individuals having $x \geq 1$ are accepted and given credit, a utility of $1,020$ units of profit may be expected, when the above settings are met.

A simulation process consisting of $1,000$ samples for 11 different values of h, ranging from 0.10 to 0.35 and for each type of estimate (non-parametric or logistic regression) was produced. The responses were generated by a procedure similar to the one described above. The non-parametric estimates of k were obtained by an iterative procedure using (5.3) and (5.6). Plots (b), (c), and (d) of Figure 5.6.1.1 illustrate the behaviour of the empirical mean squared error, variance and squared bias of the non-parametric estimates of k compared to their theoretical counterparts obtained using (5.9) and (5.10). The optimal value for h obtained using (5.12) is around 0.188, which approximately coincides with the value of h for the theoretical minimum of the mean squared error of $\hat{k}_h$. This happens even though the formula for optimal h was found maximising expected utility, so this result is quite reassuring. It can also be seen that simulated values are relatively close to theoretical ones in every case. The variance of $\hat{k}_h$ decreases as h increases and the opposite happens with its squared bias, which is what we expected from (5.9) and (5.10).

For Figure 5.6.1.2, the same simulation of 1,000 samples for each value of h and for each type of estimate is used. Plot (a) shows that the non-parametric estimate of k is positively biased and that this bias increases with h as seen before. It also shows that the logistic regression estimate coincides almost exactly with the true value of the parameter, which was expected, since the process generating the data is logistic. It is worth mentioning that only the non-parametric estimate varies with h, since obviously this is not a factor influencing the logistic regression estimation procedure. The effects of overfitting are shown in Plot (b) where the non-parametric estimate of utility exceeds the logistic regression estimate of this quantity for several values of h. Once the cross-validation correction is applied to the estimates, then utility for the non-parametric procedure decreases and goes below the one for logistic regression (Plot (c)). In this example, some values of the non-parametric cross-validated estimate of utility go a little bit above the logistic regression estimate (which is also cross-validated). This is happening because different samples where taken at each value of h to calculate the estimates. However, in simulations where the same sample was used to calculate both estimates, the non-parametric cross-validated estimate of utility always fell below the logistic regression one. It is worth noting that while the cross-validation correction makes a big difference for the non-parametric estimate, it is hardly noticeable for the logistic regression one, which is what is expected. Also, the average logistic regression estimate of utility is 1,016 units per 100,000 observations which is very close to the average maximum obtained for Plot (a) of Figure 5.6.1.1, that is, 1,020.

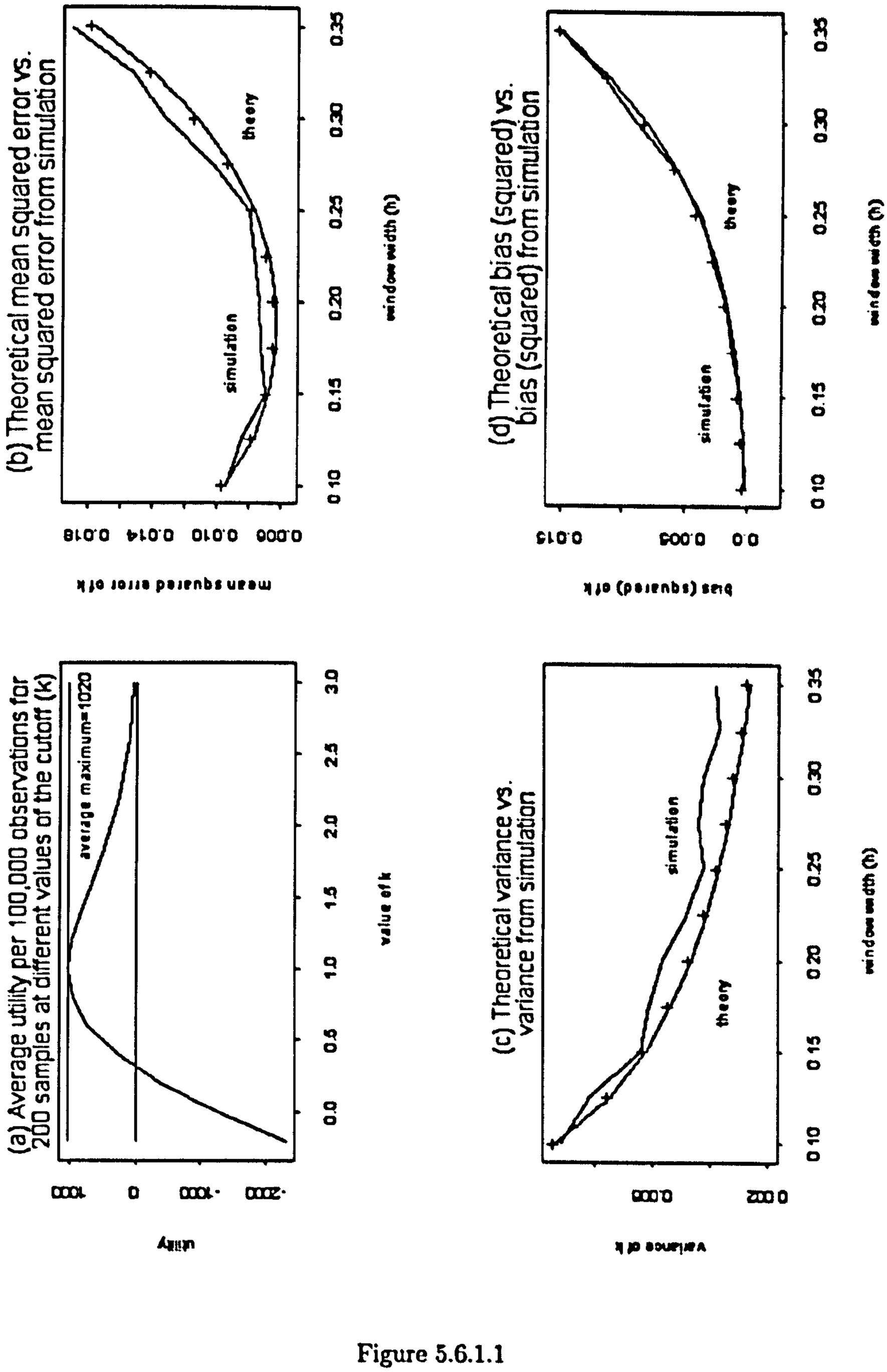
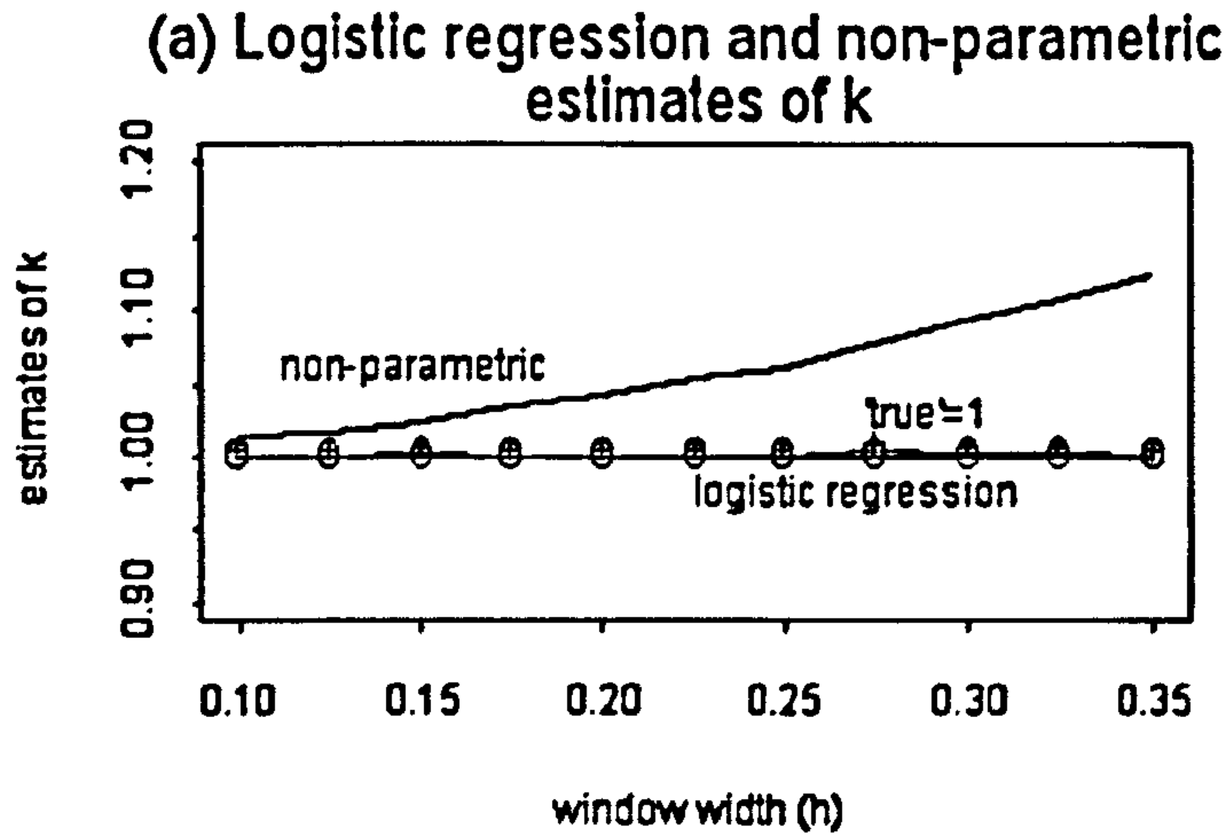
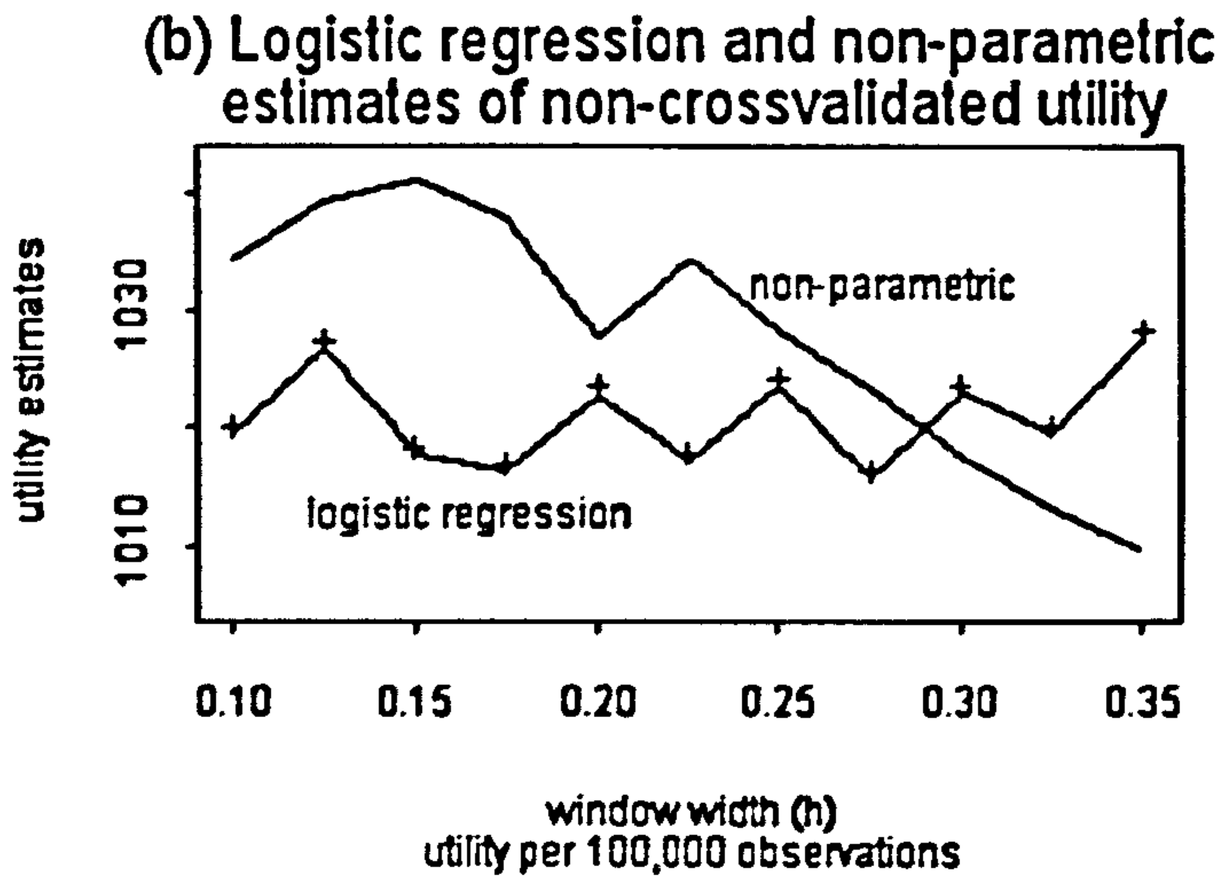

(a) Average utility per 100,000 observations for 200 samples at different values of the cutoff (k)
average maximum=1020
utility
value of k
(b) Theoretical mean squared error vs. mean squared error from simulation
theory
simulation
mean squared error of k
window width (h)
(c) Theoretical variance vs. variance from simulation
theory
simulation
variance of k
window width (h)
(d) Theoretical bias (squared) vs. bias (squared) from simulation
theory
simulation
bias (squared) of k
window width (h)

Figure 5.6.1.1


(a) Logistic regression and non-parametric estimates of k
estimates of k
0.90
1.00
1.10
1.20
non-parametric
true=1
logistic regression
0.10
0.15
0.20
0.25
0.30
0.35
window width (h)


(b) Logistic regression and non-parametric estimates of non-crossvalidated utility
utility estimates
1010
1030
non-parametric
logistic regression
0.10
0.15
0.20
0.25
0.30
0.35
window width (h)
utility per 100,000 observations

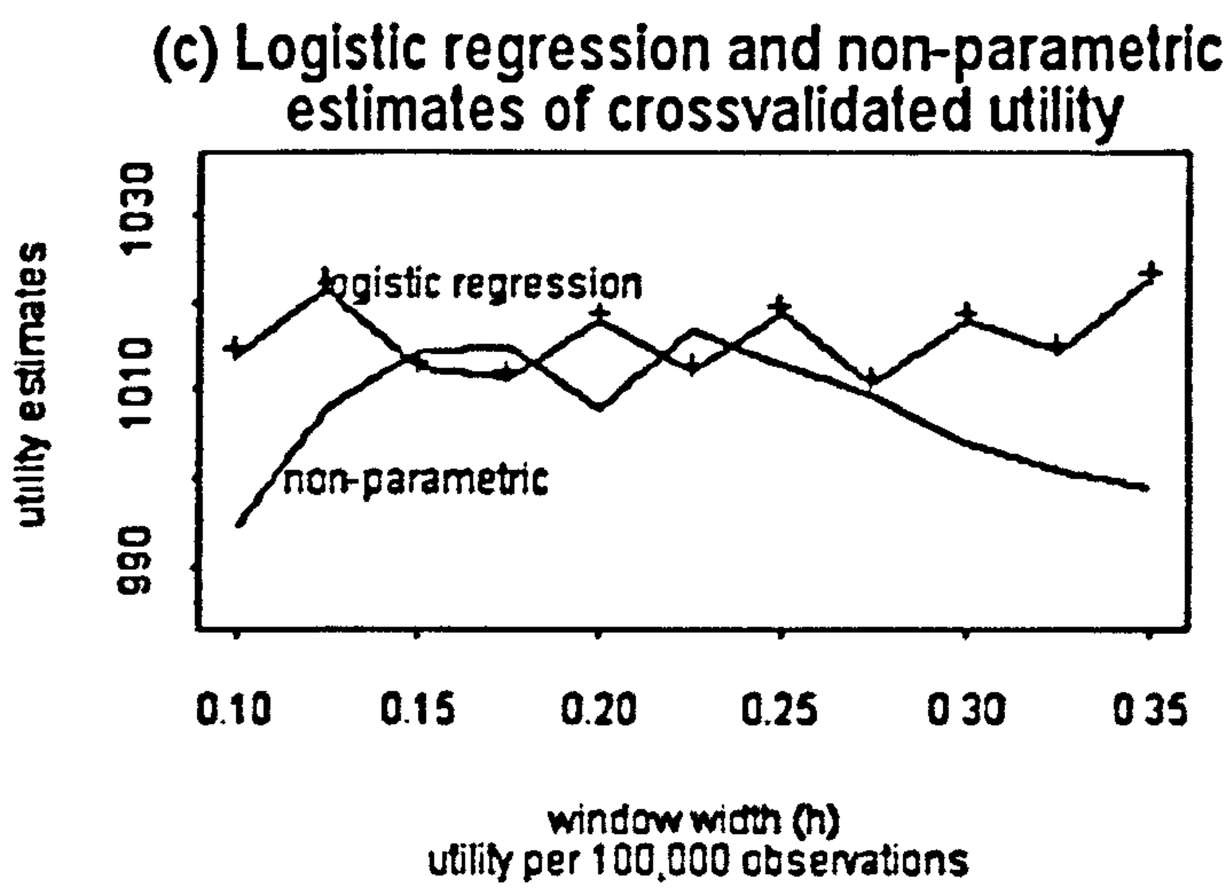

(c) Logistic regression and non-parametric estimates of crossvalidated utility
utility estimates
990
1010
1030
logistic regression
non-parametric
0.10
0.15
0.20
0.25
0 30
0 35
window width (h)
utility per 100,000 observations

Figure 5.6.1.2

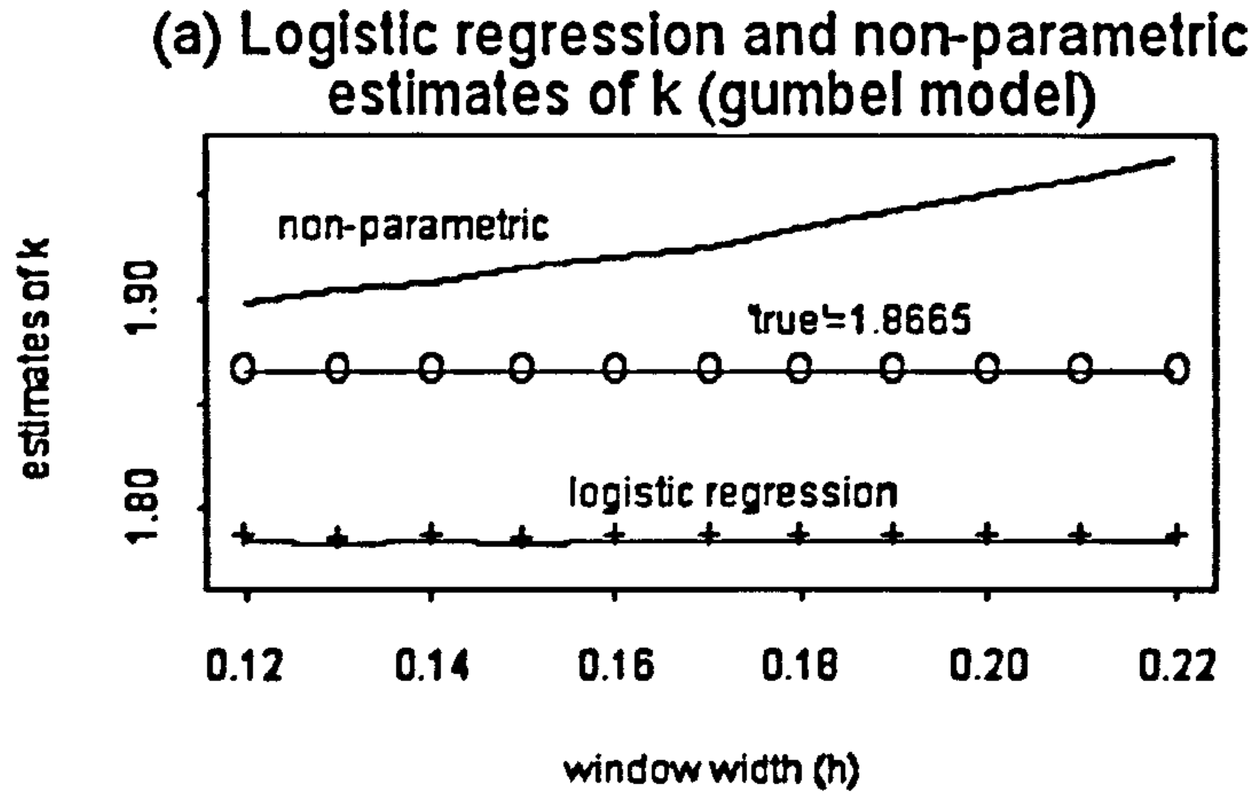
5.6.2 Example B - Gumbel Generation

A second example was explored to illustrate the behaviour of these estimates when the process generating the data is not logistic. Here, a Gumbel model was assumed for p_x. The actual model used was

$$\begin{aligned} p_x &= \exp\left(-\exp\left(-\frac{x-\alpha}{\rho}\right)\right) \\ \alpha &= 1.5 \\ \rho &= 1 \end{aligned}$$

while everything else remained the same, including the size of the simulation process. However, the range of values of h used went from 0.12 to 0.22. Responses were simulated by generating a uniform random variable in the interval $[0, 1]$ and assigning $y = 1$ if p_x, (applying the gumbel model above to each individual in the sample) was greater than this variable and $y = 0$ otherwise. Plot (a) of Figure 5.6.2.1 shows the non-parametric and logistic regression estimates of k as well as the true value of the parameter at around 1.8665. Here, the non-parametric estimate is closer to k_{true} for several values of h. Plot (b) shows cross-validated estimates of utility per 100,000 observations. It can be seen that the non-parametric estimate of utility exceeds the logistic regression estimate in almost every case, although this is more evident for values of h below 0.16. It is worth mentioning that the calculation for optimal h gives a value of 0.121. This is obtained making use of (5.12). The smallest value of h considered for the simulation was 0.12 so it may be

expected that smaller values of this quantity will result in a decrease of estimated utility. This is not shown because the estimating process becomes quite unstable for smaller values of h.


(a) Logistic regression and non-parametric estimates of k (gumbel model)
estimates of k
1.80
1.90
non-parametric
'true'=1.8665
logistic regression
0.12
0.14
0.16
0.18
0.20
0.22
window width (h)

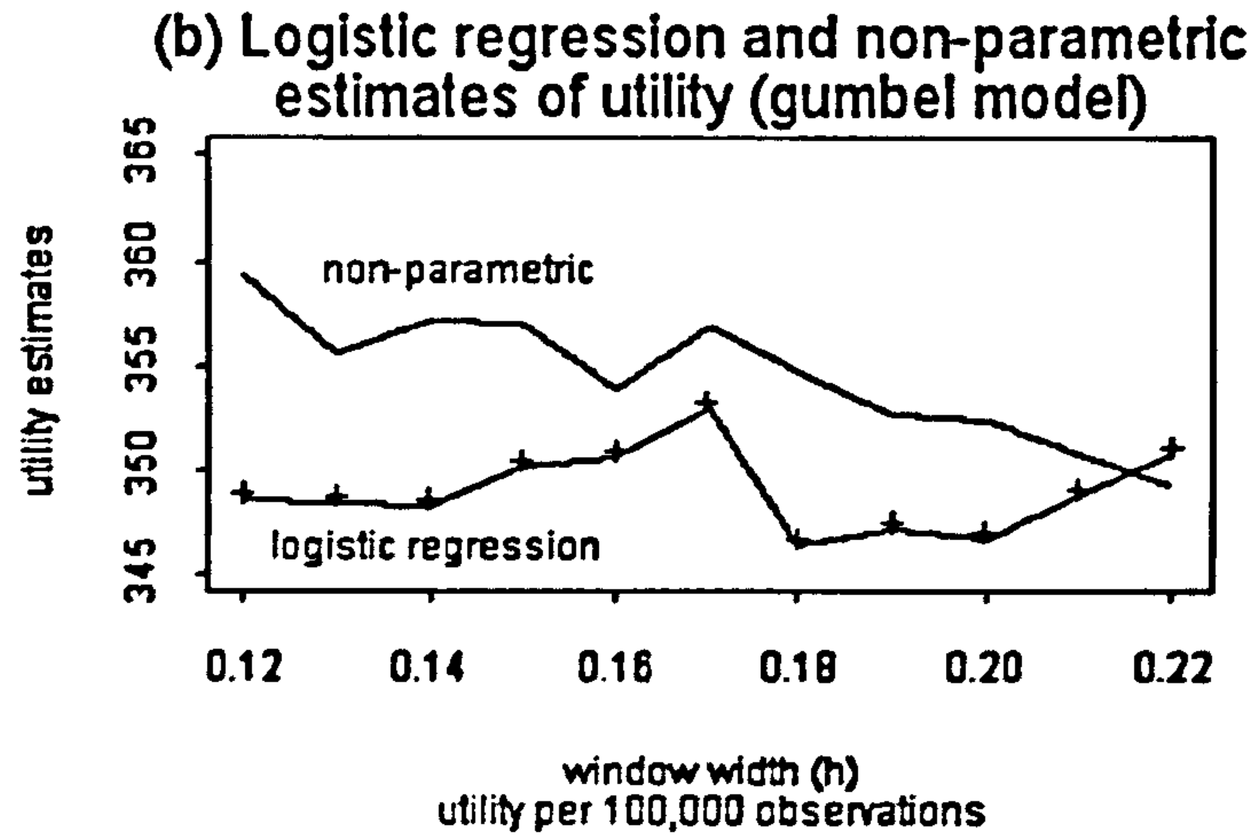

(b) Logistic regression and non-parametric estimates of utility (gumbel model)
utility estimates
345
350
355
360
365
non-parametric
logistic regression
0.12
0.14
0.16
0.18
0.20
0.22
window width (h)
utility per 100,000 observations

Figure 5.6.2.1

Chapter 6

Generalisations

6.1 Generalisation of non-parametric formulation

The idea in this section is to introduce a general weight w_i in (5.3) as any function of the covariates. Consider

$$A_{kg} = \frac{1}{nh} \sum_{i=1}^{n} w_i \, (y_i - c) \, \phi \left(\frac{x_i - k}{h} \right) \tag{6.1}$$

where w_i is any function of x.

In Figure 6.1.1 a picture of A_k using the same observations but different weights is given. The observations were generated using a logistic model to simulate responses and a normal distribution for the covariate x. It can be seen that the three curves share at least one root (in this case at $x = 1$, which is the logistic regression solution).

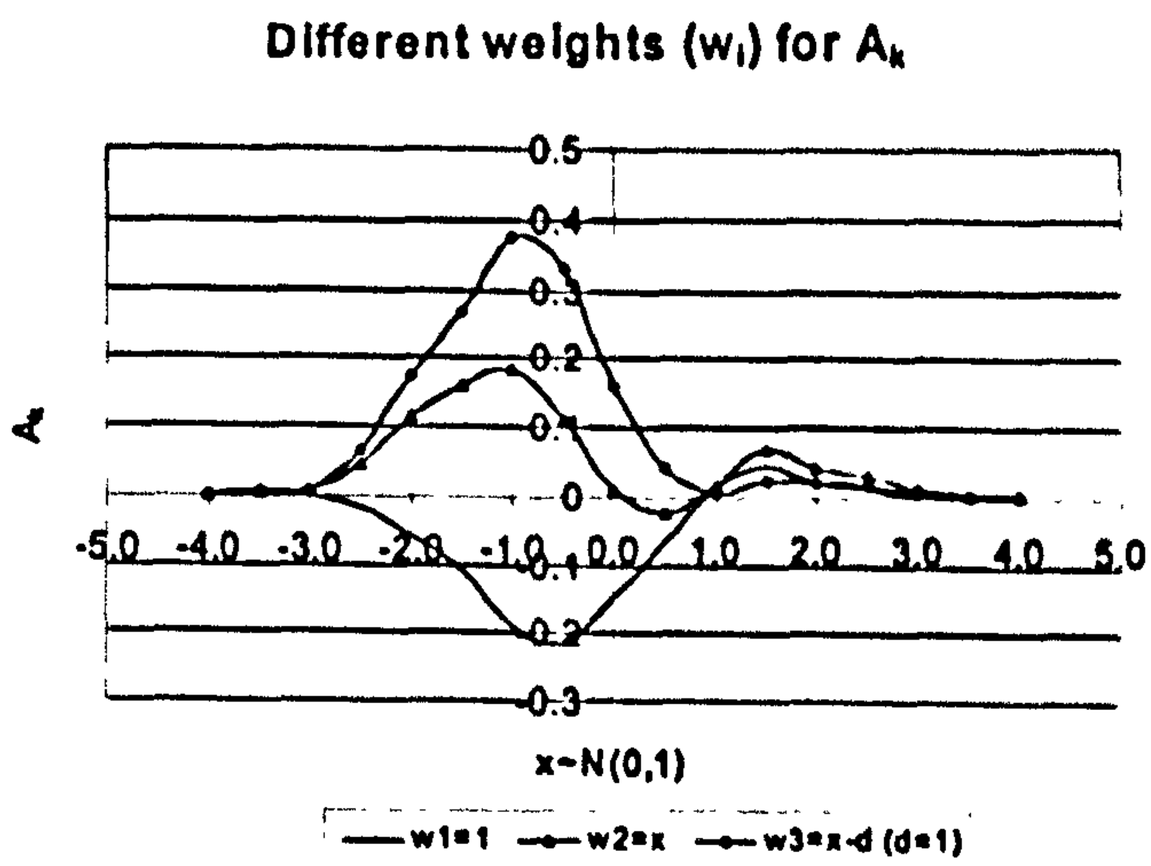


Figure 6.1.1

The procedure described in the previous section is used to obtain a generalisation for the bias and variance of k. The equations become

$$\frac{\partial}{\partial k}A_k \;=\; A'_k = \frac{1}{nh^2}\sum_{i=1}^{n} w_i\,(y_i - c)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right)$$

$$E_y\,(A'_k) \;=\; wp'f$$

$$E_y\,(A_k) \;=\; h^2 wp'f\left[f_L + \frac{w'}{w} + \frac{1}{2}\frac{p''}{p'}\right]$$

$$Var_y\,(A_k) \;=\; \frac{1}{nh}\frac{1}{\sqrt{4\pi}}wc\,(1-c)\,f$$

where w and w' denote the weight function and its first derivative, evaluated at k. The

bias and the variance become

$$Bias\left(\hat{k}_h\right) \simeq h^2\left[f_L + \frac{w'}{w} + \frac{1}{2}\frac{p''}{p'}\right] \tag{6.2}$$

$$Var\left(\hat{k}_h\right) \simeq \frac{1}{nh}\frac{1}{\sqrt{4\pi}}\frac{c(1-c)}{wf\left(p'\right)^2} \tag{6.3}$$

6.1.1 Special cases

We will consider three special cases that are of interest. These are

$$w_x = 1$$

$$w_x = x$$

$$w_x = x - d$$

In the case of $w_x = 1$ then $w' = 0$. When $w_x = x$, $w' = 1$ and $w = k$, and when $w_x = x - d$ then $w' = 1$ and $w = k - d$. The approximations for the bias and variance in each case are

$$w_x = 1 \Rightarrow Bias\left(\hat{k}_h\right) = h^2\left[f_L + \tfrac{1}{2}\tfrac{p''}{p'}\right]$$

$$Var\left(\hat{k}_h\right) = \tfrac{1}{nh}\tfrac{1}{\sqrt{4\pi}}\tfrac{c(1-c)}{(p')^2 f}$$

$$w_x = x \Rightarrow Bias\left(\hat{k}_h\right) = h^2\left[f_L + \tfrac{1}{k} + \tfrac{1}{2}\tfrac{p''}{p'}\right]$$

$$Var\left(\hat{k}_h\right) = \tfrac{1}{nh}\tfrac{1}{\sqrt{4\pi}}\tfrac{c(1-c)}{(p')^2 fk}$$

$$w_x = x - d \Rightarrow Bias\left(\hat{k}_h\right) = h^2\left[f_L + \tfrac{1}{k-d} + \tfrac{1}{2}\tfrac{p''}{p'}\right]$$

$$Var\left(\hat{k}_h\right) = \tfrac{1}{nh}\tfrac{1}{\sqrt{4\pi}}\tfrac{c(1-c)}{(p')^2 f(k-d)}$$

These approximations are just the same as before for the case when $w_x = 1$.

Another special case is when the probability of success p_x is given by the logistic model with parameters γ_0 and γ_1. In this setting the derivatives of p_x with respect to x, evaluated at $p_k = c$ are

$$p' = \gamma_1 c(1-c)$$

$$p'' = \gamma_1^2 c(1-c)(1-2c)$$

At the solution, $p_k = c$, this means that $k = \frac{\log\left(\frac{c}{1-c}\right)-\gamma_0}{\gamma_1}$. The value of γ_1 is inserted into the equations above to obtain the expressions for bias and variance of $\hat{k}_h$ in each case. When $w_x = x - d$, it is possible to find a value of d such that the bias of $\hat{k}_h$ will become zero. This might have some advantages which will be explored in more detail in § 7. However, when using the original version of A_k, given by (5.3) a relocation is not necessary, since this will be done automatically when estimating the optimal value of k.

6.2 Generalisation of cost function

6.2.1 Utility Function

In § 2 the cost of being a failure and the profit of being a success had been considered to be constants over all individuals for simplicity. However, this may not be the most realistic scenario in most applications. In this Chapter, a generalisation of the cost will be

studied, making it depend on either covariates x as defined before or on other covariates which do not influence the probability of success, say z, or both. In any case, the cost will depend on the outcome of the response variable y. That is, for a same set of covariates, the profit will in general be different to the loss if the observation turns out to be a failure instead of a success.

The cost function may be written as c_{xyz}, where subindexes indicate c always depends on y and may depend on x, on z, or on both. For the credit card application example c_{xyz} is defined as the profit obtained from an applicant if he is given credit. Then, the actual utility for a particular applicant is given by

$$I_{(0,\infty)}\left(\beta^T x - 1\right) c_{x,z,y}$$

where I is an indicator function in the interval $(0,\infty)$ with argument $\beta^T x - 1$. Then

$$\begin{aligned} E\left(U|x,y\right) &= I_{(0,\infty)}\left(\beta^T x - 1\right) E_z\left(c_{x,z,y}|x,y\right) \\ &= I_{(0,\infty)}\left(\beta^T x - 1\right) E_z\left(c_{x,z,y}|x\right) \end{aligned}$$

since z is conditionally independent of y given x. Now let $E_z\left(c_{x,z,y}|x\right) = m_x$ if $y = 1$ and $E_z\left(c_{x,z,y}|x\right) = l_x$ if $y = 0$. Then

$$E\left(U|x\right) = I_{(0,\infty)}\left(\beta^T x - 1\right)\left[m_x p_x + l_x\left(1 - p_x\right)\right]$$

So

$$E\left(U_G\right) = \int\limits_{\beta^T x>1} \left(m_x - lx\right)\left(p_x + \frac{l_x}{m_x - l_x}\right) f_x dx \tag{6.4}$$

This expression has the same form as the original utility where the costs were considered to be constant.

A special case of this expression is obtained when m_x and l_x are constants, say $m_x = a_1$ and $l_x = -a_2$, then $l_x - m_x = a_1 + a_2$ and expected utility reduces to

$$E\left(U_{G_1}\right) = \int\limits_{\beta^T x>1} \left(a_1 + a_2\right)\left(p_x - \frac{a_2}{a_1 + a_2}\right) f_x dx$$

This is the case of the original derivation, where a_1 and a_2 add up to one and $a_2 = c$.

6.2.2 Derivatives

The next step, as in the special case, is to maximise utility and therefore, first and second derivatives of utility must be obtained. From (6.4) and letting $q_x = m_x - l_x$, $r_x = \frac{l_x}{q_x}$, and $g_x = q_x\left(p_x + r_x\right) f_x$, utility is just $U = \int\limits_{\beta^T x>1} g_x dx$. So its first derivative with respect to β, apart from a constant is given by

$$U'_G = \frac{\partial}{\partial \beta} U_G \propto \int\limits_{\beta^T x=1} x g_x dx$$

Utility will be maximised for the value of β that makes $U'_G = 0$. Just as before, this

is equivalent to finding a β such that

$$E_x \left[x q_x \left(p_x + r_x \right) | \beta^T x = 1 \right] = 0$$

where previously, $q_x = 1$ and $r_x = -c$. Multiplying by β both sides indicates that utility is maximised when $E_x \left[q_x \left(p_x + r_x \right) | \beta^T x = 1 \right] = 0$, which is the same as saying that $E_x \left[q_x p_x | \beta^T x = 1 \right] = -E_x \left[q_x r_x | \beta^T x = 1 \right]$. If p_x and r_x are functions of $\beta^T x$ then they could come out of the expectations and therefore a value of β such that utility is maximised could, in principle, be obtained by solving $p = -r$. In the previous discussion for the special case, this happened when p_x was a function of $\beta^T x$, so the solution was obtained from solving $p = c$, where the absence of subindex indicates that p is evaluated precisely at $\beta^T x = 1$. For the case when p was assumed to be logistic, then the solution for β was direct.

For the univariate case, the expressions become

$$\begin{aligned} U_{Gk} &= \int_{x>k} q_x \left(p_x + r_x \right) f_x dx \\ U'_{Gk} &= -q_k \left(p_k + r_k \right) \end{aligned}$$

So $U'_k = 0 \Leftrightarrow p_k = -r_k$ where the solution is the root in k of $p_k + r_k = 0$.

Going back to the multivariate case, the second derivative, again apart from a

constant, is

$$U''_G = \frac{\partial}{\partial \beta^T} U' \propto - \int_{\beta^T x = 1} \left[I + x \left(\frac{\partial \log g_x}{\partial x} \right)^T \right] \beta x^T g_x dx$$

And since $g_x = q_x (p_x + r_x) f_x$, then

$$\frac{\partial \log g_x}{\partial x} = \frac{q'_x}{q_x} + \frac{p'_x + r'_x}{p_x + r_x} + \frac{f'_x}{f_x}$$

where the dashed versions of the functions denote their derivatives with respect to x. So, at the solution for U'_G

$$U''_G \propto - \int_{\beta^T x = 1} x \left(\frac{q'_x}{q_x} + \frac{p'_x + r'_x}{p_x + r_x} + \frac{f'_x}{f_x} \right)^T \beta x^T q_x (p_x + r_x) f_x dx \tag{6.5}$$

Before, when the costs were constants, $r_x = -c$ so $r'_x = 0$ and $q_x = 1$ so $U'' \propto - \int_{\beta^T x = 1} \beta^T p'_x x x^T f_x dx$.

As before, formula (6.5) will simplify when both p_x and r_x are functions of $\beta^T x$. In this case, the quantities involving $\frac{q'_x}{q_x}$ and $\frac{f'_x}{f_x}$ will vanish from the integral at the solution for β. Also, if a particular model is assumed for p_x and r_x, then derivatives of these functions could be obtained in order to simplify (6.5)even more and, therefore, get

$$U''_{G1} \propto - (p' + r')^T \beta \int_{\beta^T x = 1} x x^T q_x f_x dx \tag{6.6}$$

6.2.3 Estimation

Suppose a sample of values of x and y is obtained. Let $E_x\left(c_{x,x,y}|x\right) = m_x$ if $y_i = 1$ and $E_x\left(c_{x,x,y}|x\right) = l_x$ if $y_i = 0$. To estimate k in the univariate case using the non-parametric procedure the functions are the following

$$A_{Gk} = \frac{1}{nh}\sum_{i=1}^{n} q_{x_i}\left(y_i + r_{x_i}\right)\phi\left(\frac{x_i - k}{h}\right) \tag{6.7}$$

The first derivative of A_{Gk} with respect to k is given by

$$A'_{Gk} = \frac{1}{nh^2}\sum_{i=1}^{n} q_{x_i}\left(y_i + r_{x_i}\right)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right)$$

and its expectation over y

$$E_y\left(A'_{Gk}\right) = \frac{1}{nh^2}\sum_{i=1}^{n} q_{x_i}\left(p_{x_i} + r_{x_i}\right)\left(\frac{x_i - k}{h}\right)\phi\left(\frac{x_i - k}{h}\right) \tag{6.8}$$

Expanding $p_x + r_x$ around $x = k$, where k is the solution to A_{Gk} gives

$$p_x + r_x \simeq \left(p_k + r_k\right) + \left(p'_k + r'_k\right)\left(x - k\right) = \left(p'_k + r'_k\right)\left(x - k\right)$$

since $p_k = -r_k$. Here p'_k and r'_k denote the first derivatives of p_x and r_x with respect to x, evaluated at k. So, an approximation to (6.8) is given by

$$E_y\left(A'_{Gk}\right) \simeq \frac{1}{nh}\left(p'_k + r'_k\right)\sum_{i=1}^{n} q_{x_i}\left(\frac{x_i - k}{h}\right)^2 \phi\left(\frac{x_i - k}{h}\right) \tag{6.9}$$

In the above expression, an idea of the form of p_x and r_x would be useful in order to obtain their derivatives. For estimation purposes, formulas (6.7) and (6.9) would be used iteratively to obtain $\hat{k}_h$.

For the multivariate case, the above expressions become

$$A_{G\beta} = \frac{1}{nh}\sum_{i=1}^{n} x_i q_{x_i}\left(y_i + r_{x_i}\right)\phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{6.10}$$

$$A'_{G\beta} = -\frac{1}{nh^2}\sum_{i=1}^{n} x_i x_i^T q_{x_i}\left(y_i + r_{x_i}\right)\left(\frac{\beta^T x_i - 1}{h}\right)\phi\left(\frac{\beta^T x_i - 1}{h}\right)$$

where $A'_{G\beta}$ is the derivative of $A_{G\beta}$ with respect to β. Expectation of $A'_{G\beta}$ with respect to y is given by

$$E_y\left(A'_{G\beta}\right) = -\frac{1}{nh^2}\sum_{i=1}^{n} x_i x_i^T q_{x_i}\left(p_{x_i} + r_{x_i}\right)\left(\frac{\beta^T x_i - 1}{h}\right)\phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{6.11}$$

and expanding $p_{x_i} + r_{x_i}$ around $\beta^T x_i = 1$ gives

$$p_x + r_x \simeq (p + r) + (p' + r')\left(\beta^T x - 1\right) = (p' + r')\left(\beta^T x - 1\right)$$

since $p = -r$ at the solution. Here, p' and r' are derivatives of p_x and r_x with respect to $\beta^T x$, evaluated at $\beta^T x = 1$, assuming both are functions of $\beta^T x$. This gives the following approximation for $E\left(A'_{G\beta}\right)$

$$E\left(A'_{G\beta}\right) \simeq -\frac{1}{nh}\left(p' + r'\right)\sum_{i=1}^{n} x_i x_i^T q_{x_i}\left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right) \tag{6.12}$$

Again, an idea of the form of p_x and r_x would be useful in order to obtain their derivatives. For estimation purposes, formulas (6.10) and (6.12) would be used iteratively to obtain $\hat{\beta}_h$.

6.2.4 Special Cases

Special cases involve the use of a known model for p_x and now r_x as well. For example, if for the univariate case the following was true

$$\begin{aligned} l_x &= a + bx \\ m_x &= c + dx \\ q_x &= (c-a) + (d-b)\,x \\ r_x &= \frac{a + bx}{(c-a) + (d-b)\,x} \end{aligned}$$

and also

$$p_x = \frac{\exp\left(\gamma_0 + \gamma_1 x\right)}{1 + \exp\left(\gamma_0 + \gamma_1 x\right)},$$

then k would be the solution (in x) to

$$\frac{\exp(\gamma_0 + \gamma_1 x)}{1 + \exp(\gamma_0 + \gamma_1 x)} = -\frac{a + bx}{(c - a) + (d - b)\, x}$$

and derivatives of p_x and r_x with respect to x

$$\begin{aligned} p'_x &= \gamma p_x (1 - p_x) \\ r'_x &= \frac{bc - ad}{[(c - a) + (d - b)\, x]^2} \end{aligned}$$

These derivatives evaluated at k would be introduced in (6.9) for estimation purposes.

A similar situation would arise for the multivariate case.

6.2.5 An application

A simple example related to screening for a disease is presented as an illustration for the ideas discussed in the present Chapter. The data used for the example was taken from the Office for National Statistics website (www.statistics.gov.uk). Two tables were considered

- Prevalence of diagnosed diabetes: by sex and age, 1998: Social Trends 33
- Table 3: Interim revised population estimates; England and Wales, 1992 – 2000

Both tables refer to the population in England and Wales. The table for prevalence of diagnosed diabetes by sex and age for 1998 is presented below

Age	Males	Females
0 – 15	1.3	1.5
16 – 24	3.7	3.6
25 – 34	5.7	4.7
35 – 44	12.3	8.9
45 – 54	25.8	17.4
55 – 64	55.9	38.8
65 – 74	84.7	64.0
75 – 84	87.2	65.5
85+	76.3	60.4

Table 6.2.5.1

The rates are given by 1,000 population. To obtain an estimate of the number of individuals with diagnosed diabetes for the year 2000, the rates of prevalence of diagnosed diabetes in 1998 were applied to the population projections for each age group for 2000. In general, when dealing with this type of information, males and females are treated separately. This is because of the differences due to gender in human biology. However, for illustration purposes, in this example the information of males and females is aggregated and handled as one population. The next table provides the estimated number of individuals with diagnosed diabetes and the estimate of the total number of individuals in each age group (the figures represent thousands).

Age	Total Population	Diagnosed Diabetes
0 – 15	10,540	15
16 – 24	5,582	20
25 – 34	7,703	40
35 – 44	7,655	81
45 – 54	6,874	148
55 – 64	5,441	257
65 – 74	4,366	322
75 – 84	2,903	215
85+	1,007	65
Total	52,071	1,163

Table 6.2.5.2

For purposes of the example suppose the total population corresponds to individuals (healthy and diseased) where diabetes has not been diagnosed yet. One way to make the example more realistic would be to estimate, through a survey or otherwise, the rates of undiagnosed diabetes in the population by age group. This would provide an approximate number of individuals who are not diagnosed with the illness and the prevalence of the disease among this group, by age. It will also be considered that the prevalence of diagnosed diabetes coincides with the true prevalence of the disease in the undiagnosed population.

Suppose the NHS is willing to screen for diabetes from a certain age in order to diminish the costs of treating the illness once it is detected. It is known that diabetes is an illness that gradually deteriorates the body over time, especially when not treated. Therefore, early diagnosis represents a gain in terms of costs of treatment as well as quality of life and well being for the patient. Also, the gain an early detection brings could depend on age, being greater for older patients. Another issue to consider is that screening itself is a procedure that involves a certain cost. Suppose that the test used for screening is 100% reliable, being always positive for individuals with the disease and always negative when the disease is not present. This is also a simplification of the real world, where false positive and false negative rates are usually greater than zero. Suppose also that once the disease is detected, the overall cost of treatment from this moment until the death of the patient could be determined by the patient's age at diagnosis.

The costs for the example will be the following

- Cost of test for screening: -1 unit
- Gain for detecting diabetes: 1.2 units times the age at detection

Therefore the cost functions, using the notation presented above are

$$l_x = E(c_x|x) = -1 \text{ if } y = 0 \text{ (not diseased)}$$

$$m_x = E(c_x|x) = -1 + 1.2x \text{ if } y = 1 \text{ (diseased)}$$

$$r_x = \frac{l_x}{m_x - l_x} = -\frac{1}{1.2x}$$

where x is the age of the patient. The objective now is to find the optimal age to start screening for diabetes in the undiagnosed population, depending on these costs. All the figures used represent thousands in the population.

The first step is to calculate the actual overall cost or profit by age and the utility obtained starting screening at each age. A plot of this utility function is given in Figure 6.2.1. It can be seen that the maximum utility is achieved at around 50 years of age, where a profit of around 61,000 units is obtained.

For the example, data were generated following the approximate distribution of the population and the disease. In total, a data set of 52,059 individuals for which diabetes has not been diagnosed was obtained. From these, it was estimated that 1,166 individuals would actually have the disease. The procedure to obtain the logistic regression estimate is also based on a numerical approximation, since the optimal value of k (age at which screening should start) is given by the root of

$$\frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 k)}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 k)} = -r_k = \frac{1}{1.2k}$$

where $\hat{\gamma}_0 \simeq -6.265$ and $\hat{\gamma}_1 \simeq 0.0488$ are obtained from the data. The root of this function is around 46.39. So the solution given by logistic regression indicates that screening should start at around 46 years and 5 months of age ($k = 46.4$). This provides an estimate of utility of around 60,862 units. For the non-parametric procedure, the

functions used are

$$A_{Gke} = \frac{1}{52,059h} \sum_{i=1}^{52,059} 1.2x \left(y_i - \frac{1}{1.2x} \right) \phi \left(\frac{x_i - k}{h} \right)$$

$$E_y \left(A'_{Gk} \right) \simeq \frac{1}{52,059h} \left(\hat{\gamma} p_k \left(1 - p_k \right) + \frac{1}{1.2k^2} \right) \sum_{i=1}^{52,059} 1.2x \left(\frac{x_i - k}{h} \right)^2 \phi \left(\frac{x_i - k}{h} \right)$$

where x is age, p_k represents the logistic approximation for p_x evaluated at k, and $\hat{\gamma} \simeq 0.0488$ as before. It was found that a value of $h = 0.25$ works fairly well for the data. The approximate value for k obtained using these functions is around 44.56 years of age, which is a slightly younger age to start screening than the logistic regression estimate for this value. The estimate of utility if screening is started at 44 years and 6 months of age is around 61,124 units, which is slightly larger than the figure obtained using a logistic regression model for estimation.

This example gives an idea of how the theory presented here may be applied to a real situation and provides encouraging results as to the benefits that might be obtained.

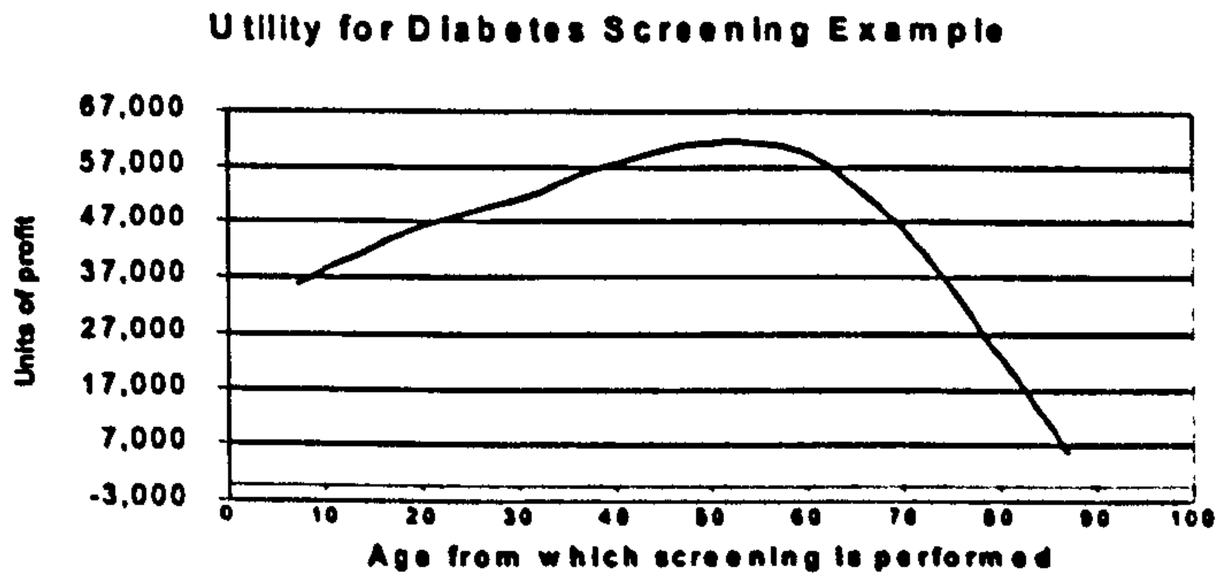


Figure 6.2.1

Chapter 7

Relocation of Covariates

7.1 Transformation of x

In this chapter, a location shift of the covariates will be explored. This is considered to be relevant since is was observed that shifts have an important effect on the bias of $\hat{\beta}_h$.

Take $x^* = x - d$ and suppose that p_x is logistic. Then

$$\begin{aligned}
\Pr(y=1|x) &= p_x = \frac{\exp(\gamma_0 + \gamma^T x)}{1+\exp(\gamma_0 + \gamma^T x)} \\
&= \frac{\exp(\gamma_0 + \gamma^T (x^* + d))}{1+\exp(\gamma_0 + \gamma^T (x^* + d))} \\
&= \frac{\exp(\gamma_0 + \gamma^T d + \gamma^T x^*)}{1+\exp(\gamma_0 + \gamma^T d + \gamma^T x^*)} \\
&= \frac{\exp(\gamma_0^* + \gamma^{*T} x^*)}{1+\exp(\gamma_0^* + \gamma^{*T} x^*)} = p_x^*
\end{aligned}$$

where $\gamma_0^* = \gamma_0 + \gamma^T d$ and $\gamma^* = \gamma$. Also, if $b^* = \log \frac{c}{1-c} - \gamma_0^*$ and $\gamma = b\beta$ then

$$b^* = b - \gamma^T d = b - b\beta^T d = b\left(1 - \beta^T d\right) = ba$$

where $a = 1 - \beta^T d$. Now, if $p_x^* = c$ which is the same as saying that $\log \frac{c}{1-c} = \gamma_0^* + \gamma^{*T} x$, then $\gamma^{*T} x = \log \frac{c}{1-c} - \gamma_0^* = b - \gamma^T d$. So

$$\frac{\gamma^{*T} x}{b - \gamma^T d} = 1$$

and

$$\beta^* = \frac{\gamma}{b - \gamma^T d} = \frac{\gamma}{ba} = \frac{\beta}{a}$$

Also if

$$\beta^{*T} x^* = \frac{\beta^T (x - d)}{a} = s$$

then

$$\beta^T x = as + \beta^T d = as + 1 - a = 1 + a(s - 1)$$

We also have

$$\mu_s^* = E_{x^*}\left[x^* | \beta^{*T} x^* = s\right] = E_x\left[(x - d) | \beta^T x = 1 + a(s - 1)\right]$$

So

$$\mu_s^* = \mu_{1+a(s-1)} - d$$

and therefore,

$$\mu_s'^* = a\mu'_{1+a(s-1)}$$

Then $\mu^* = \mu - d$ and $\mu^{*\prime} = a\mu'$. Also

$$\begin{aligned} V_s^* &= E_{x^*}\left[x^* x^{*T} | \beta^{*T} x^* = s\right] \\ V^* &= V - d\mu^T - \mu d^T + dd^T \end{aligned}$$

where $V = E_x\left[xx^T | \beta^T x = 1\right]$ and

$$\frac{\beta^{*T} x^* - 1}{h^*} = \frac{\frac{\beta^T (x-d)}{a} - 1}{h^*} = \frac{\beta^T (x-d) - a}{ah^*} = \frac{\beta^T x - 1}{ah^*}$$

so $ah^* = h$ or $h^* = h/a$. Finally

$$g_s^* = \frac{d}{ds} \Pr\left(\beta^{*T} x^* < s\right) = \frac{d}{ds} \Pr\left(\beta^T x < 1 + a\left(s-1\right)\right) = a g_{1+a(s-1)}$$

and

$$g_s^{*\prime} = a^2 g_{1+a(s-1)}$$

with $g^* = ag$ and $g^{*\prime} = a^2 g'$. So

$$g_L^* = \frac{\partial}{\partial s} \log g^* \,|_{s=1} = \frac{g^{*\prime}}{g^*} = a\frac{g'}{g} = a\frac{\partial}{\partial s} \log g \,|_{s=1} = a g_L \tag{7.1}$$

From (3.28) the bias of $\hat{\beta}_h^*$ is given by

$$Bias\left(\hat{\beta}_h^*\right) \simeq h^{*2}V^{*-1}\left[\mu^*\left(g_L^* + \frac{1}{2}b^*\left(1-2c\right)\right) + \mu^{*\prime}\right]$$

and substituting the above values this is

$$\begin{aligned} Bias\left(\hat{\beta}_h^*\right) &\simeq \frac{h^2}{a^2}\left(V - d\mu^T - \mu d^T + dd^T\right)^{-1} \times \\ &\quad \left[(\mu - d)\left(ag_L + \frac{1}{2}ab\left(1-2c\right)\right) + a\mu'\right] \\ &= \frac{h^2}{a}\left(V - d\mu^T - \mu d^T + dd^T\right)^{-1} \times \\ &\quad \left[(\mu - d)\left(g_L + \frac{1}{2}b\left(1-2c\right)\right) + \mu'\right] \end{aligned}$$

where $a = 1 - \beta^T d$. From this formula it can be seen that if $d = 0$, we obtain the previous expression for the bias of $\hat{\beta}_h$. Also a particular choice of d can bring the approximation of the bias to zero. Therefore in the Taylor expansion for the approximation an extra term may be added, increasing the order to h^4. Ideally, this would provide a larger window and hence more observations having larger weights when estimating $\hat{\beta}_h$.

The value of d that makes the bias zero is

$$d_{opt} = \mu + \frac{1}{g_L + \frac{1}{2}b\left(1-2c\right)}\mu' \tag{7.2}$$

An important aspect is that the optimal d is obtained using the distribution of x and also p_x. In practice, usually these two elements will be unknown. One possibility is to

take the logistic estimate of p_x and a distribution for x that reasonably fits the data to obtain an estimate of d_{opt}.

7.2 Approximation for $Bias\left(\hat{\beta}_h\right)$ under special relocation

If $x^* = x - d_{opt}$, where d_{opt} is as given by (7.2), the approximation for the bias of $\hat{\beta}_h$ requires an extra term in the Taylor expansion used. This is because the original approximation of the bias will become zero under this particular transformation. In (3.22), considering an extra term in the Taylor expansion gives

$$E_y\left(A^*_{\beta_h}\right) \simeq h^{*2} g^* \left[M^{*\prime} g^*_L + \frac{1}{2} M^{*\prime\prime} + h^{*2}\left(\frac{1}{4} M^{*\prime\prime} \frac{g^{*\prime\prime}}{g^*} + \frac{1}{6} M^{*\prime\prime\prime} g^*_L\right)\right]$$

but when $d = d_{opt}$ then $M^{*\prime} g^*_L + \frac{1}{2} M^{*\prime\prime} = 0$ so

$$E_y\left(A^*_{\beta_h}\right) \simeq h^{*4} g^* \left[\frac{1}{4} M^{*\prime\prime} \frac{g^{*\prime\prime}}{g^*} + \frac{1}{6} M^{*\prime\prime\prime} g^*_L\right]$$

Since $E_y\left(A^{*\prime}_{\beta_h}\right)$ given by (3.23) evaluated at x^* remains the same, the bias is now given by

$$Bias\left(\hat{\beta}^*_h\right) \simeq h^{*4} \left[N^* g^*_L + N^{*\prime}\right]^{-1} \left[\frac{1}{4} M^{*\prime\prime} \frac{g^{*\prime\prime}}{g^*} + \frac{1}{6} M^{*\prime\prime\prime} g^*_L\right] \tag{7.3}$$

for the most general case. The expression for the variance remains the same as before. The value of h which produces a maximum expected utility is now given by

$$h^*_{opt} = \left\{ \frac{tr\left(R^* U^{*\prime\prime}_\beta\right)}{8n\delta^{*T}\left(B^*\right)^{-1T} U^{*\prime\prime}_\beta \left(B^*\right)^{-1}\delta^*} \right\}^{1/9} \propto n^{-1/9} \tag{7.4}$$

where

$$\begin{aligned} B^* &= N^* g^*_L + N^{*\prime} \\ \delta^* &= \frac{1}{4}\frac{g^{*\prime\prime}}{g^*} M^{*\prime\prime} + \frac{1}{6} M^{*\prime\prime\prime} g^*_L \\ W^* &= E_{x^*}\left[x^* x^{*T} p_{x^*}\left(1 - p_{x^*}\right) \mid \beta^{*T} x^* = 1\right] \\ R^* &= \frac{1}{g^*}\frac{1}{\sqrt{4\pi}}\left[B^*\right]^{-1} W^* \left[B^*\right]^{-1} \end{aligned}$$

where the stars indicate the functions are evaluated at x^*.

7.3 Special Cases

When the probability of success is based on the score, that is $p^*_x = p^*_s$, then the bias is expressed as follows

$$Bias\left(\hat{\beta}^*_h\right) \simeq h^{*4} V^{*-1} \alpha^*_3 \left[\mu^* \alpha^*_1 + \mu^{*\prime} \alpha^*_2 + \mu^{*\prime\prime\prime}\right] \tag{7.5}$$

where

$$\begin{aligned}
\alpha_3^* &= \frac{1}{2} g_L^* \\
\alpha_2^* &= \frac{g^{*\prime\prime}}{g^{*\prime}} + \frac{p^{*\prime\prime}}{p^{*\prime}} \\
\alpha_1^* &= \frac{1}{2}\frac{p^{*\prime\prime}}{p^{*\prime}}\frac{g^{*\prime\prime}}{g^{*\prime}} + \frac{1}{3}\frac{p^{*\prime\prime\prime}}{p^{*\prime}} \\
V^* &= E_{x^*}\left[x^* x^{*T} \mid \beta^{*T} x^* = 1\right]
\end{aligned}$$

When p_s^* is logistic, the appropriate expressions for $p^{*\prime}$, $p^{*\prime\prime}$, and $p^{*\prime\prime\prime}$ can be substituted above. These substitutions will lead to the optimal value of h^*.

$$h_{opt}^* = \left\{ \frac{c(1-c)r}{8n\sqrt{4\pi} g^* \delta_l^{*T} V^{*-1} \delta_l^*} \right\}^{1/9} \tag{7.6}$$

where $\delta_l^* = \alpha_3^* \left[\mu^* \alpha_1^* + \mu^{*\prime} \alpha_2^* + \mu^{*\prime\prime}\right]$ and α_3^*, α_2^*, and α_1^* are defined as before. Also $\mu^* = \mu - d$, $\mu^{*\prime} = a\mu'$, and $\mu^{*\prime\prime} = a^2\mu''$. Under the logistic assumption for p_s^*, the appropriate values of the $\alpha's$ may be inserted in (7.6).

7.4 Normal distribution for covariates x

Suppose $x \sim N(m, \Sigma)$ so $\beta^T x \sim N\left(\beta^T m, \beta^T \Sigma \beta\right)$. Considering the results from § 3.1.3 we have that $g_L = -\frac{1}{2}\frac{\left(1-\beta^T m\right)}{\beta^T \Sigma \beta}$ and μ and μ' are given by (3.33) and (3.34) respectively.

Suppose also that p_x is logistic with $\text{logit}(p_x) = \gamma_0 + \gamma^T x$. Then the optimal value of d

for a location shift is given by

$$
\begin{aligned}
d_{optn} &= m + \not{\Sigma}\beta \left(\frac{1}{\beta^T \not{\Sigma} \beta} \right) (1 - \beta^T m) + \\
&\qquad \frac{1}{-\frac{1}{2} \frac{(1-\beta^T m)}{\beta^T \Sigma \beta} + \frac{1}{2} b (1 - 2c)} \frac{\not{\Sigma}\beta}{\beta^T \not{\Sigma} \beta} \\
d_{optn} &= m + \left[(1 - \beta^T m) + \frac{1}{-\frac{1}{2} \frac{(1-\beta^T m)}{\beta^T \not{\Sigma} \beta} + \frac{1}{2} b (1 - 2c)} \right] \frac{\not{\Sigma}\beta}{\beta^T \not{\Sigma} \beta} \qquad (7.7)
\end{aligned}
$$

where $b = \log\left(\frac{c}{1-c}\right) - \gamma_0$. Now suppose $x^* = x - d_{optn}$. Then $x^* \sim N(m^*, \not{\Sigma})$ where $m^* = m - \left(\mu + \frac{1}{g_L + (1/2)b(1-2c)} \mu'\right)$ and $\beta^{*T} x^* \sim N(\beta^{*T} m^*, \beta^{*T} \not{\Sigma} \beta^*)$. If X^* is a vector formed by x^* and $\beta^{*T} x^*$ so that $X^* = \left[\begin{smallmatrix} x^* \\ \beta^{*T} x^* \end{smallmatrix}\right]$ then the distribution of X^* is given by

$$
X^* \sim N \left[\begin{pmatrix} m^* \\ \beta^{*T} m^* \end{pmatrix}; \begin{pmatrix} \not{\Sigma} & \not{\Sigma}\beta^* \\ \beta^{*T} \not{\Sigma} & \beta^{*T} \not{\Sigma} \beta^* \end{pmatrix} \right]
$$

From this distribution, the expressions for (3.31), (3.33), (3.34), and for the density of $\beta^{*T} x^*$ may be obtained following the procedure of § 3.1.3.

7.5 Example

A very simple example will be examined in this section. This example will provide a graphic illustration of what happens with the bias of $\hat{\beta}_h$ when a location shift takes place. It will also help illustrate the advantage of introducing a shift which is optimal, such as the one given by (7.2) in terms of the mean squared error and the optimal value

for h. The example will consider only two covariates for simplicity, but the results when working with a different number of covariates are completely analogous.

Please refer to Example A of § 4.2.1. The same setting will be used here. The first thing to consider will be the calculation of the optimal value of d, say d_{optn}. The values for μ, μ', and g_L are given by

$$\mu = \begin{bmatrix} 0.6923 \\ 0.6154 \end{bmatrix} = \mu'; g_L = -0.2564$$

where μ and μ' coincide because $m = \binom{0}{0}$. Also, $b = 1$ and $c = 0.5$ and $a = 1 - \beta^T d = -1.95$. Therefore, the optimal d is given by

$$d_{optn} = \begin{bmatrix} -0.6577 \\ -0.5846 \end{bmatrix}$$

When using x, the optimal value for h obtained by (5.12) is around 0.22. The optimal value for h when $x^* = x - d_{optn}$ and using (7.6) is around 0.21. These two values are not very different because the optimal h is based on maximum utility and not on minimal bias. However, having unbiased estimates is always preferable. In this example, if a simulation is carried out, following the procedure described in § 4.2.1, the utility obtained using the optimal relocation of the covariates is very similar to the one using the raw covariates. Figures 7.5.1 and 7.5.2 show how the bias and mean squared error for each parameter are reduced. In this case, the optimal range of h in terms of mean squared error changes from around $(0.20, 0.25)$ to around $(0.30, 0.35)$. This means a larger window width could

be taken, which would imply a larger number of observations would receive larger weights and, therefore, an improvement in the estimation process.

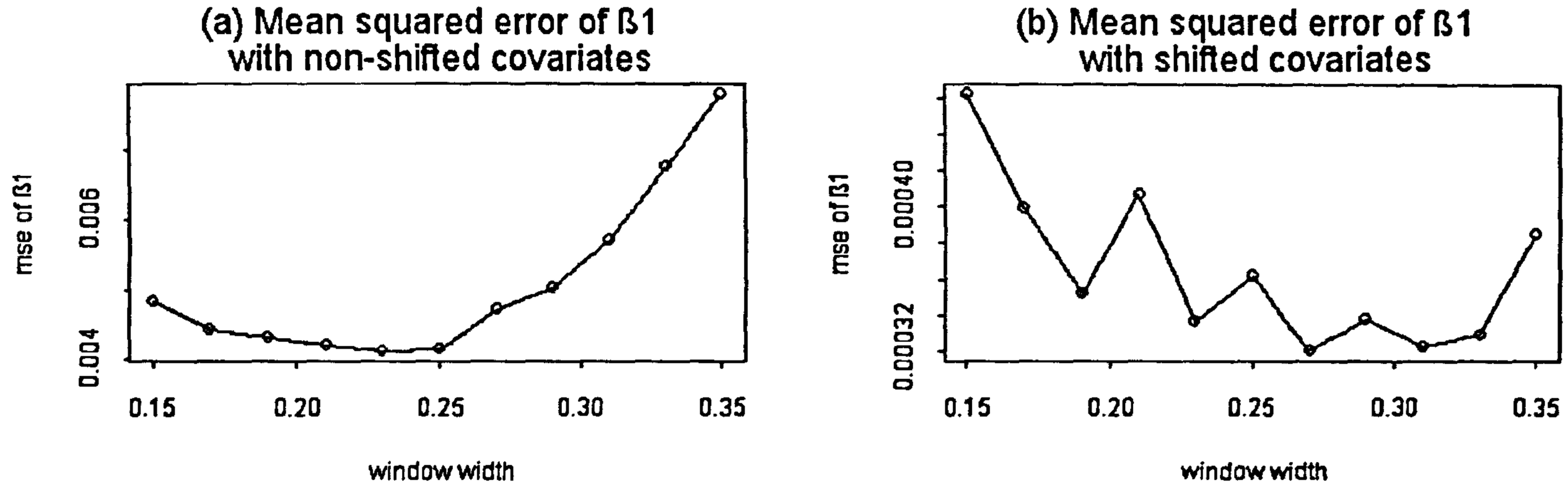
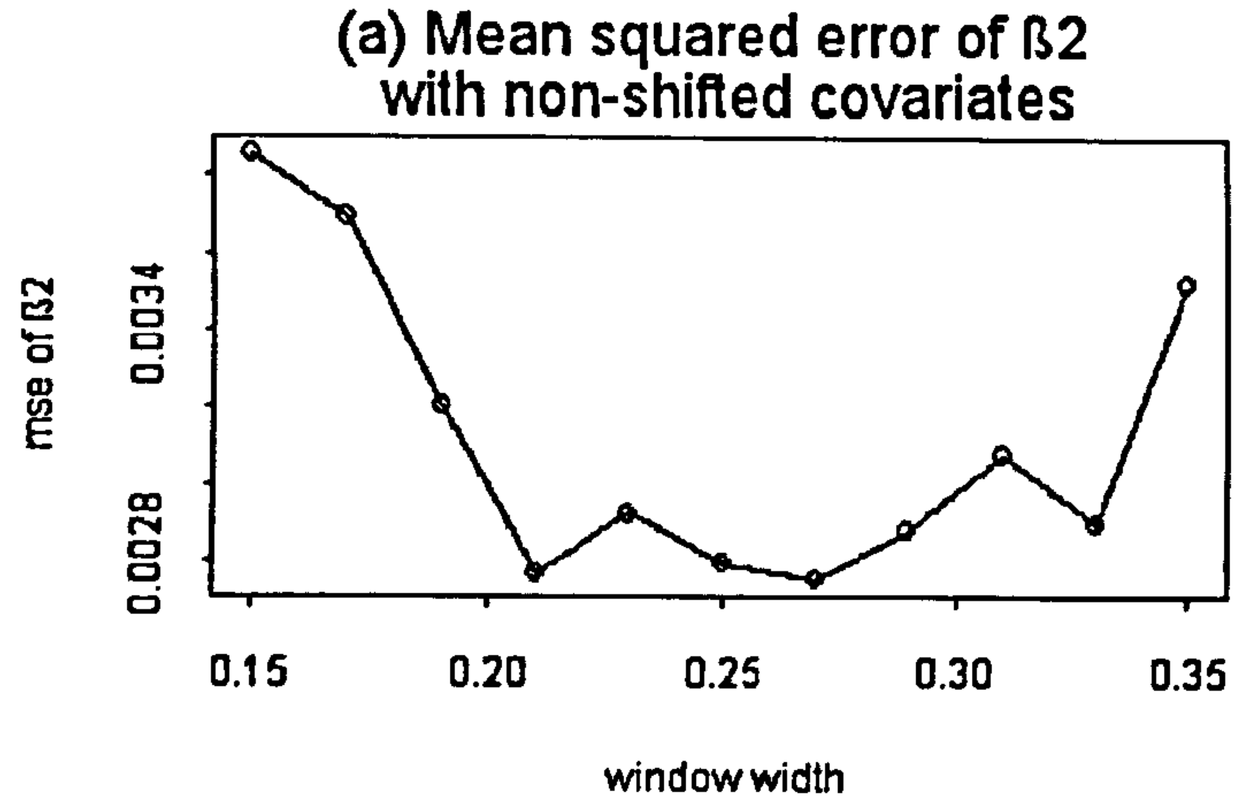
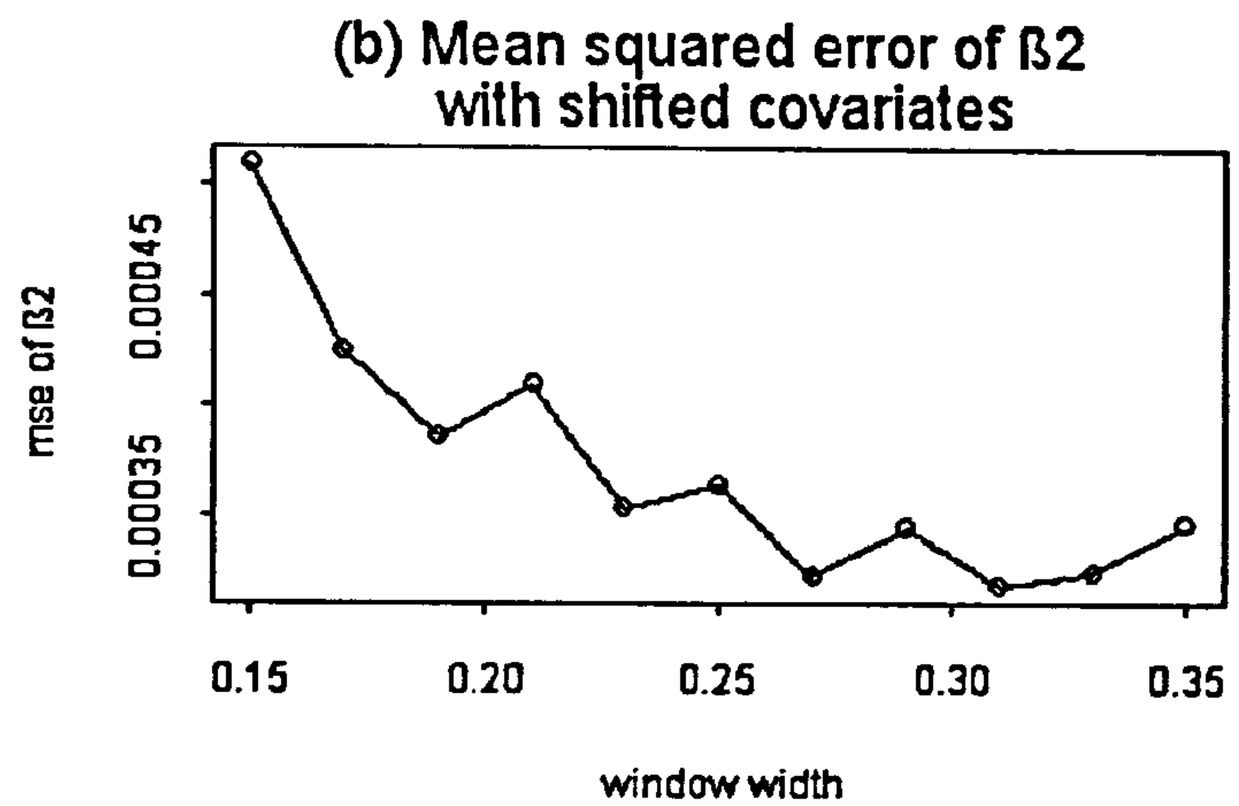
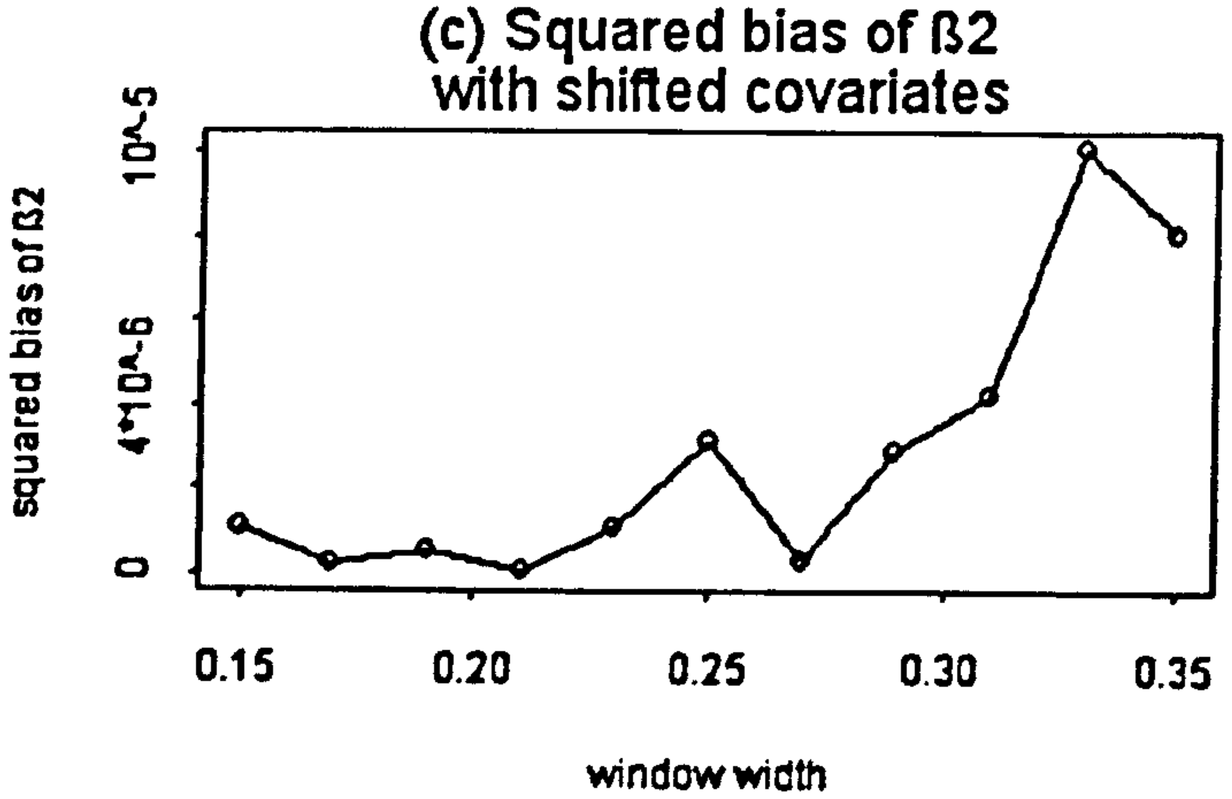

(a) Mean squared error of ß1
with non-shifted covariates
mse of ß1
0.004
0.006
0.15
0.20
0.25
0.30
0.35
window width
(b) Mean squared error of ß1
with shifted covariates
mse of ß1
0.00032
0.00040
0.15
0.20
0.25
0.30
0.35
window width
(c) Squared bias of ß1
with shifted covariates
squared bias of ß1
0.0
0.00002
0.00005
0.15
0.20
0.25
0.30
0.35
window width
(d) Comparison of mse(ß1)
for shifted and non-shifted covariates
mse of ß1
0.0
0.004
non-shifted covariates
shifted covariates
0.15
0.20
0.25
0.30
0.35
window width

Figure 7.5.1


(a) Mean squared error of ß2
with non-shifted covariates
mse of ß2
0.0028
0.0034
0.15
0.20
0.25
0.30
0.35
window width


(b) Mean squared error of ß2
with shifted covariates
mse of ß2
0.00035
0.00045
0.15
0.20
0.25
0.30
0.35
window width


(c) Squared bias of ß2
with shifted covariates
squared bias of ß2
0
4*10^-6
10^-5
0.15
0.20
0.25
0.30
0.35
window width

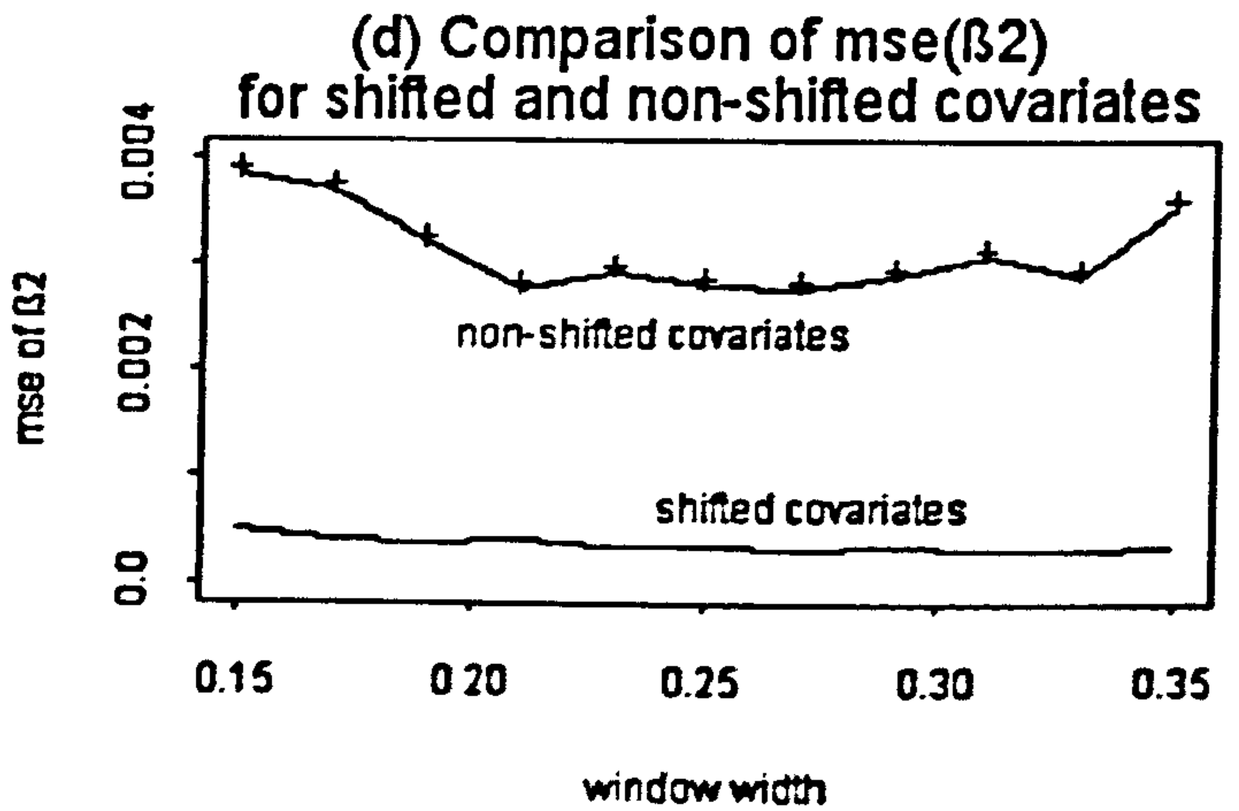

(d) Comparison of mse(ß2)
for shifted and non-shifted covariates
mse of ß2
0.0
0.002
0.004
non-shifted covariates
shifted covariates
0.15
0.20
0.25
0.30
0.35
window width

Figure 7.5.2

Chapter 8

Case Study

8.1 Description of data

The case study presented here relates to the motivation of the thesis. The data base used for illustration purposes corresponds to credit card application information from one of the largest banks in Mexico. Since the information is confidential, only the results will be shown, not the individual cases. In this section, a thorough description of the data base will be given.

The original data set had to be cleaned in order to obtain a base where all the information was consistent and the fields for every covariate contained logical information for all observations. The observations where missing fields were found were deleted from the data base. We believe that this process was done without introducing any bias since the missing fields are not due to a particular process or situation. The final data

base consists of 24,057 observations. These are all entries from applicants for credit card during the year going from the beginning of July 1997 to end of June 1998 who were accepted as clients and granted a credit card. This implies that a previous selection process was carried out in order to identify those applicants that would ultimately become "good" customers. Application information, along with credit bureau records was captured. This is what makes up the covariate data base. To obtain the response variable, a definition of a bad (in other words, defaulted) account had to be considered. It is common to consider an account as defaulted when the client fails to repay his or her loan (at least the minimum payment in credit cards) for three months in a row or more during a certain amount of time known as performance period. This is the amount of time the behaviour of the account is observed in order to determine if it is good or bad. The performance period is usually one year considered from the moment the credit is granted. The observations in the data base were classified in good and bad using this definition of default. Furthermore, accounts with an undesirable status, but that were not defaulted under this definition were also classified as defaults. Since the base is composed of accounts opened during one whole year, each account is referred to its own performance period.

Approximately 11% of the observations in the data base were classified as defaulted. Application information and credit bureau records were obtained for everyone. However, it is important to mention that there exists a data base containing all the rejected applicants corresponding to this time period. These observations do not have a performance

indicator. That is, we do not know if they would have been defaults or not. The idea behind the thesis tries to tackle exactly this point. That is, we are trying to obtain the linear predictor and cutoff point under this predictor (who is accepted and who is rejected) that will maximise the utility of the bank. Since we do not have performance information about rejected applicants, we will consider, for illustration purposes, that the accepted individuals represent the ones that were rejected. This, in the sense that a rejected observation having same covariate information as an accepted one would produce a similar performance. This is just the same as finding an expansion factor associated to every accepted individual. The weight is given by the number of applicants from the population (with same covariate information) that this accepted individual represents in the actual accepted base. This is not the best way to go, but what can be done with this kind of information. The ideal situation would consist of a controlled trial process in which the bank would accept every applicant and observe performance after a year. The linear predictor and cutoff point obtained for this kind of information would represent exactly the population from where the trial base was obtained. Even though this would be a better way to start producing application models, it represents a tough decision for a bank, since a big loss could be expected coming from the increased default rate and greater outstanding balances. Some banks tackle this problem by creating controlled trial cells in specific populations of interest (students, low incomes, pensioners) where they test the behaviour of these groups when given a credit card.

The approach taken for this case study was to assign an expansion factor to each

accepted applicant in the base. This expansion factor was calculated considering the information given by the covariates in the complete data base (including rejected applicants). A table showing the counts for the weighted and unweighted versions of the base, split by default indicator, is given below.

	Defaulted	Not defaulted	Total
Unweighted	2,645	21,412	24,057
%	10.99	89.01	100
Weighted	5,554	41,173	46,727
%	11.89	88.11	100

Table 8.1.1

The information contained in the covariates is classified in three big categories. The first category involves classification information such as account or client number and the credit score assigned by the previous selection process. The second category involves credit bureau information such as number of revolving credit lines (e.g. credit cards), months since the oldest account was opened (seniority handling loans), and satisfactory references (number of loans in a "good" status). Finally, the third category refers to demographic information. This involves economic dependants, gender, residential status (owner, renting, living with relatives, etc.), kind of employment or job, marital status (single, married, widower, free union, etc.), date of birth, income, time at present home, and time at present employment. The natural logarithm of income was used instead of income and the covariates with reference to time in months were transformed to time in years. Also, date of birth was transformed to age.

8.2 Exploratory analysis

The exploratory analysis of the data included a preliminary bivariate analysis where each covariate was tested against the response (default). Since default is a categorical variable, the first idea was to perform a chi-squared independence test to observe association between default and each of the covariates. However, this is useful for categorical covariates, where there is a small number of categories and each cell has a reasonable count. For almost all numerical covariates this is not an option, unless bins are created to concentrate the information. This process was done in order to reduce the information and perform the chi-squared tests.

More graphical types of analysis included plotting each covariate against the default percent in each level of the covariate. These plots also include an idea of the amount of observations in each of these levels. This is helpful in order to identify trends in default rates among categories of numerical covariates and their impact. The plots are presented in the Appendix to this Chapter. Another type of plot, which may be used both for numerical and for categorical covariates is one where each level of the covariate is plotted against the ratio of good customers in that level and the total of good customers (G/TG). On the same plot, the line of the ratio between bad customers in that level and total bad customers (B/TB) may be drawn. This gives an idea of those categories where the proportion of "good" customers is significantly different from the proportion of bad customers, if there are any. The Appendix to this Chapter shows these plots for all the numerical covariates that were selected using the independence chi-squared tests

described above. Another type of plot is the ROC (Receiver Operating Characteristic) curve where a score s relating default to the covariate is calculated (using regression for instance). The curve is formed by the plot of $\Pr(s \geq u|y=1)$ against $\Pr(s \geq u|y=0)$ as u ranges over all possible values of the score s. The more this curve is away from the 45° line, the more useful the covariate is to discriminate between good and bad customers and, in this sense, the more the covariate is related to default. A quantity called discriminating power may be calculated for these curves. Depending on the relationship between the covariate and default (positive or negative) the ROC curve may go above or below the 45° line. The power is calculated as the area under the ROC curve for curves going above the 45° line and one minus this area for curves going below the 45° line. The larger the power, the better the covariate is in terms of discriminating between good and bad customers. Again, in the Appendix to this Chapter these curves are shown for the selected numerical covariates. Also, histograms for numerical covariates are given.

8.3 Optimal window h and convergence issues

Assuming a logistic model, an approximate optimal value for h was calculated using formula (3.30). An idea of the model selection process followed will be given in § 8.4, but for purposes of estimating the optimal window width, suppose the covariates that will be introduced in the model are already known. For this calculation, a normality assumption for the covariates was used, even though this is clearly not the case.

Under a logistic assumption for p_x, say $p_x = \frac{\exp(\gamma_0+\gamma^T x)}{1+\exp(\gamma_0+\gamma^T x)}$ where $\gamma = (\gamma_1, \gamma_2, ..., \gamma_r)$

and $r = 12$, estimates for (γ_0, γ) are given below (the categorical covariate "employment" was re-defined in dummies)

variable name	parameter	value
intercept	γ_0	−0.06468854
economic dependants	γ_1	0.1023967
satisfactory references	γ_2	−0.008975956
age	γ_3	0.04087302
log income	γ_4	0.3083291
years since oldest account opened	γ_5	0.639101
dummy 1 - housewife, student, pensioner	γ_6	0.1434613
dummy 2 - shopkeeper / trader	γ_7	0.220196
dummy 3 - businessman	γ_8	0.05002197
dummy 4 - teacher	γ_9	0.2192308
dummy 5 - worker / technician	γ_{10}	0.2671407
dummy 6 - executive position	γ_{11}	−0.9318136
dummy 7 - secretary / office worker	γ_{12}	0.3487437

Table 8.3.1

The value of c used for the analysis is 0.85. This was selected because it is the level at which the bank feels safe. It was calculated with the formula for expected default frequency the bank is already using for new accounts. This is given by

$$EDF = \frac{1}{1 + bedf * slope^{CS}}$$

where $bedf = 0.02131$, $slope = 1.0276$ and CS is the credit score (score assigned by a previous model, indicating if the account was to be accepted or rejected). The minimal previous credit score is around 205, which corresponds to an expected default frequency of 0.15 and therefore a probability of success of 0.85 at the cutoff point, which corresponds to the value of c.

Using these values and assumptions, an optimal value for h was calculated. This value is approximately 0.02, which is quite small. In fact, the smallest value that could be used without incurring in convergence problems during the non-parametric estimation process was $h = 0.15$. This value of h produces estimates for the $\beta's$ which are quite close to the estimates given by logistic regression. However, it has to be pointed out that whenever using a real data set, problems of this kind will generally arise, especially with a large number of covariates. The good news is that the non-parametric procedure proved to be quite effective at giving a bigger utility than the logistic regression procedure. This means that, on one hand, the assumptions used to calculate the optimal h are not quite satisfactory, and in this sense it confirms that perhaps the logistic model is not appropriate for this kind of data. The other one is that, even though the value of h used in the estimation process is quite larger than this "optimal" value, the results are still better than those obtained for the logistic regression in terms of utility. It also needs to be said that even if the utility using the non-parametric procedure is somewhat larger than the one obtained using logistic regression, it is also true that the difference between both is not very big.

8.4 Model selection, estimation, and comparison between non-parametric and logistic regression estimates

Following parsimony and based on the exploratory analysis of § 8.2, the covariates selected for modelling were $x = (x_1, ..., x_{12})$

- x_1 - economic dependants - demographic (numerical)
- x_2 - satisfactory references - credit bureau (numerical)
- x_3 - age - demographic (numerical)
- x_4 - natural logarithm of income - demographic (numerical)
- x_5 - years since oldest account opened - credit bureau (numerical)
- x_6 to x_{12} - type of employment - demographic (categorical)

Since type of employment is a categorical covariate with eight categories, seven dummy variables were created to introduce this covariate into the models. Even though default also showed a dependence to residential status and gender, these covariates were not introduced in the models because the difference in deviance resulting from introducing them was very small. That is, they did not improve the model in a significant way and thus were not included. The expansion factor mentioned in § 8.1 was included as a

weight in both estimations as fex_i. The logistic regression model was estimated using the generalised linear model routine included in *Splus*, using the model

$$p_x = \frac{e^{\gamma_0+\gamma^T x}}{1+e^{\gamma_0+\gamma^T x}}$$

where γ_0 is the intercept and $\gamma = (\gamma_1, \gamma_2, ..., \gamma_{12})$ is the parameter for x as mentioned above. The logistic regression estimates for β were calculated as $\hat{\beta} = \frac{\hat{\gamma}}{\log\left(\frac{c}{1-c}\right)-\hat{\gamma}_0}$ where $c = 0.85$ and $\hat{\gamma}_0$ and $\hat{\gamma}$ are the estimates of the logistic regression model. Using (3.10) and (3.14) from §3.2, the functions used to obtain the estimates through the non-parametric procedure are

$$A_\beta = \frac{1}{nh}\sum_{i=1}^{n} x_i\,(y_i - c)\,\phi\left(\frac{\beta^T x_i - 1}{h}\right) fex_i$$

$$E\,(A'_\beta) = \frac{bc\,(1-c)}{nh}\sum_{i=1}^{n} x_i x_i^T \left(\frac{\beta^T x_i - 1}{h}\right)^2 \phi\left(\frac{\beta^T x_i - 1}{h}\right) fex$$

where $b = \log\left(\frac{c}{1-c}\right) - \hat{\gamma}_0$ (the logistic regression estimate), $n = 24057$, $h = 0.15$, and fex is the expansion factor again. Using $\hat{\beta}_h \simeq \beta - \left[E\,(A_\beta)'\right]^{-1} A_\beta$ and following an iterative process, estimates for β were obtained.

A comparison of the results obtained for the estimates of the linear predictor using logistic regression and the non-parametric procedure follows. This comparison is made in terms of the estimates of the parameters and in terms of the utility obtained using each procedure. The next table gives the estimates of the $\beta's$ for each covariate using both, along with the differences between them and the utility in each case.

covariate	non-parametric	logistic reg	difference
economic dependants	−0.02567	−0.02426	−0.00141
satisfactory references	0.038877	0.038402	0.00047
age	−0.00366	−0.00337	−0.0003
log income	0.116241	0.115634	0.00061
years since oldest acct opened	0.016304	0.015329	0.00097
d1-housewife,stud.,pensioner	0.256148	0.239686	0.01646
d2-shopkeeper / trader	0.063148	0.053803	0.00934
d3-businessman	0.089078	0.082581	0.0065
d4-teacher	0.011126	0.01876	−0.00763
d5-worker / technician	0.087708	0.082219	0.00549
d6-executive position	0.094557	0.100187	−0.00563
d7-secretary / office worker	0.14262	0.130791	0.01183
utility (non-crossvalidated)	1,743	1,719	25
utility (crossvalidated)	1,720	1,707	13

Table 8.4.1

As mentioned before, the differences between the estimates of the parameters are not very big and this is reflected in the utility estimates, where differences are not big either. With the exception of economic dependants and age, all other covariates are positively correlated with success.

The performance of both models is almost exactly identical. For the purpose of comparing both, an ROC curve was created for each model and the power calculated. These plots may be seen in Figures 8.4.1.*a* and 8.4.1.*b*.

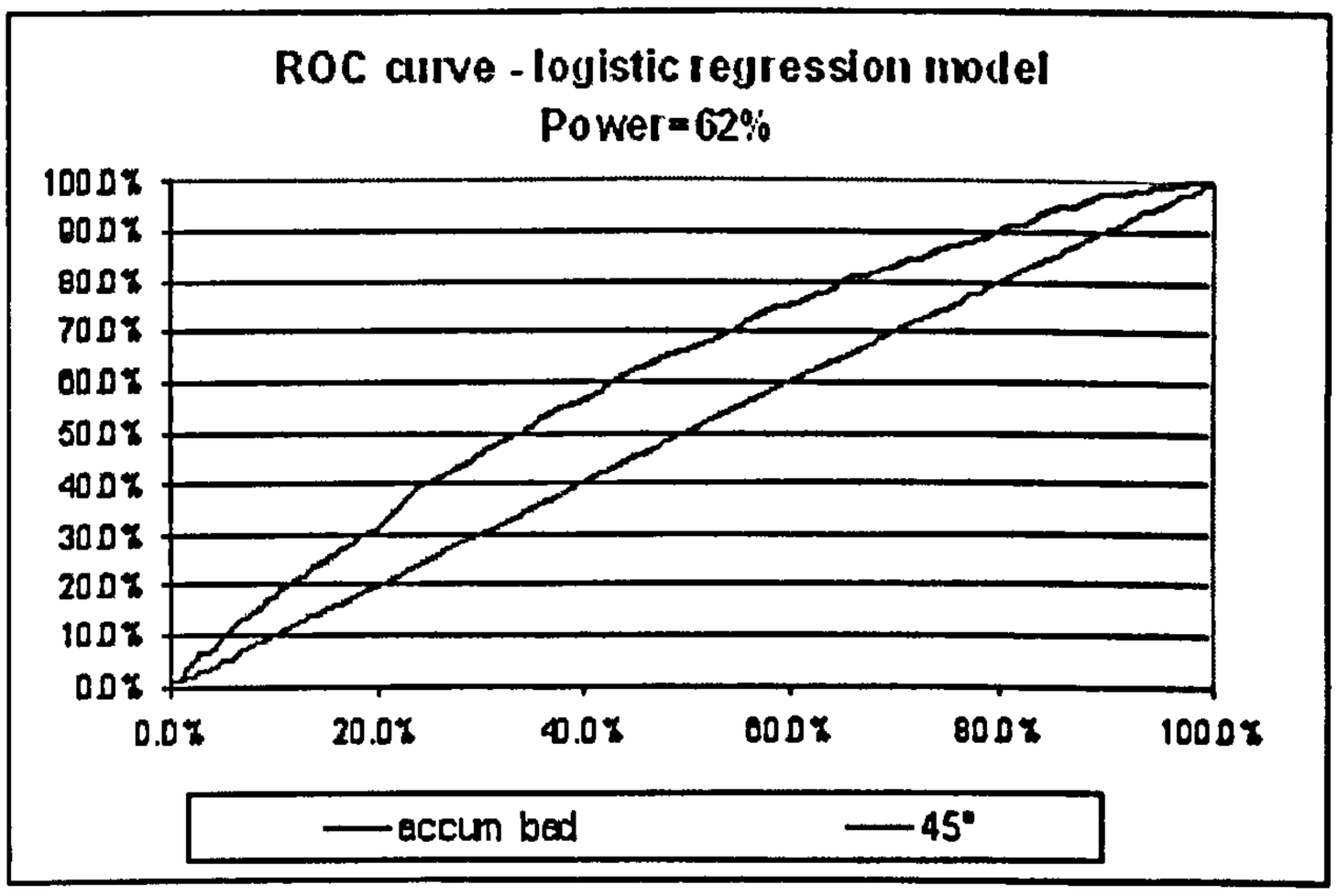


Figure 8.4.1.a.

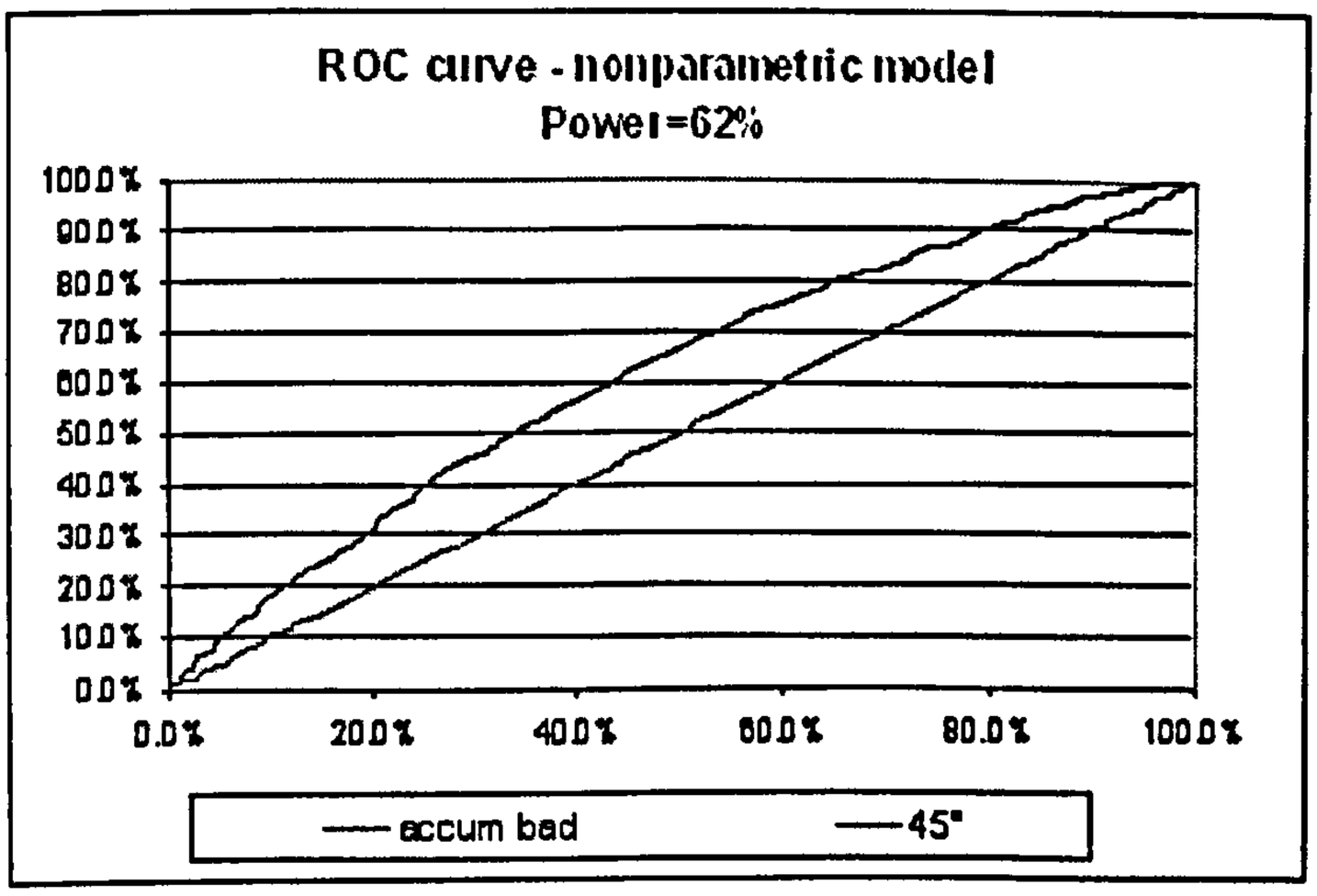
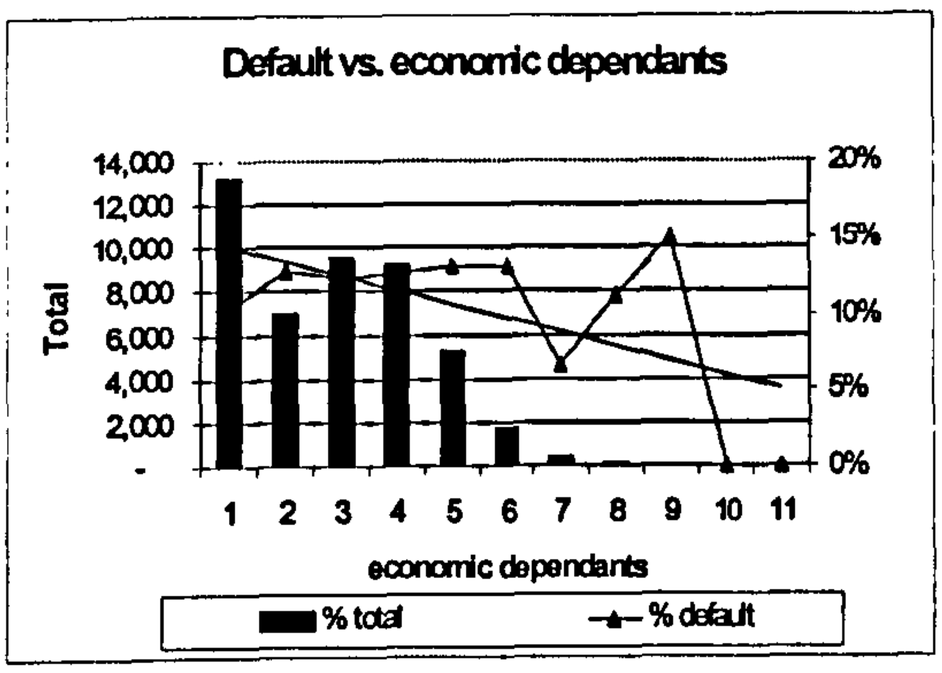
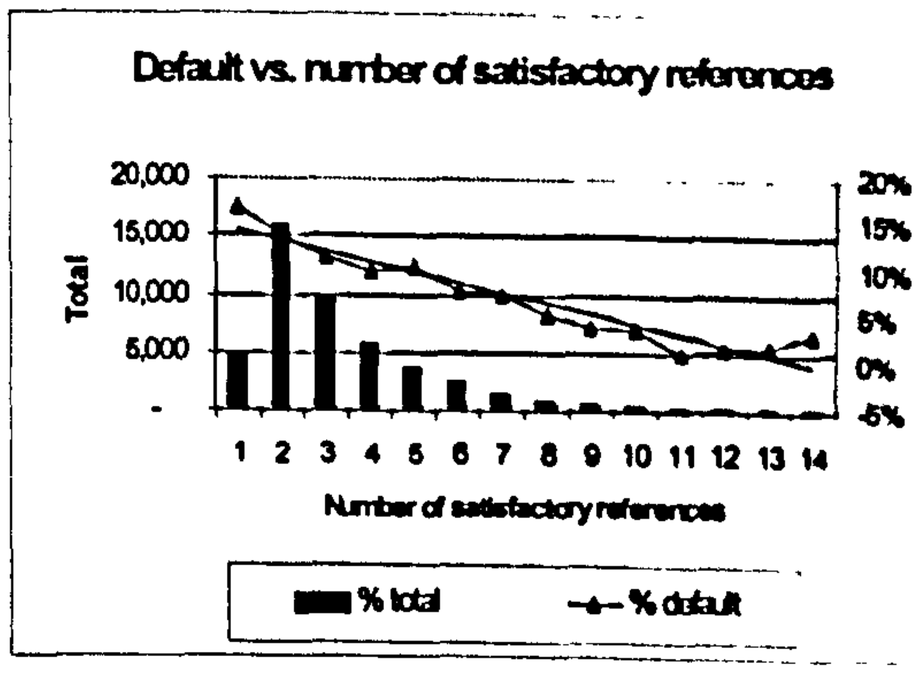
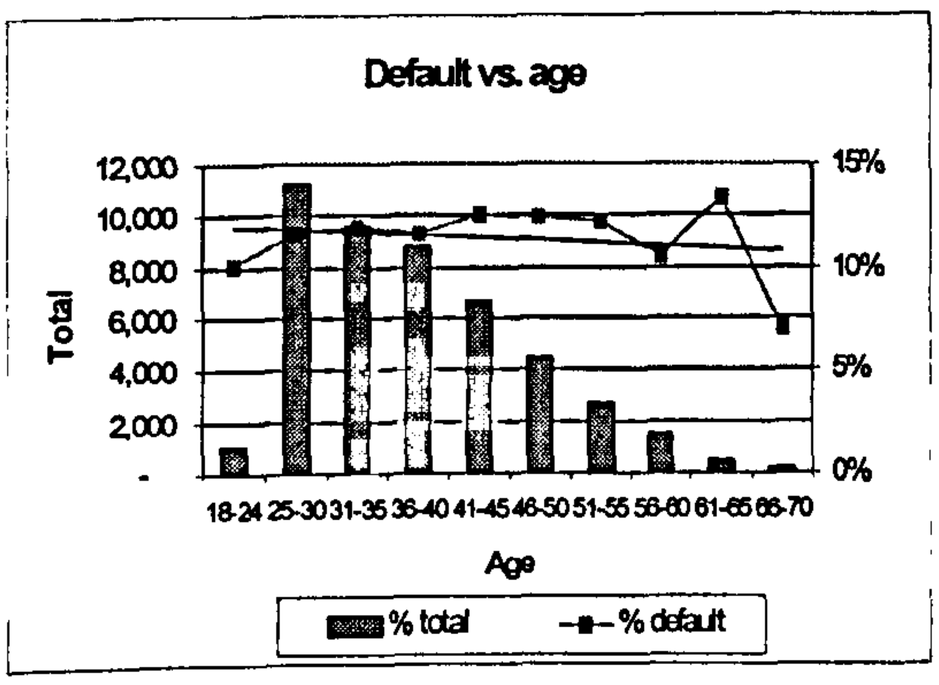


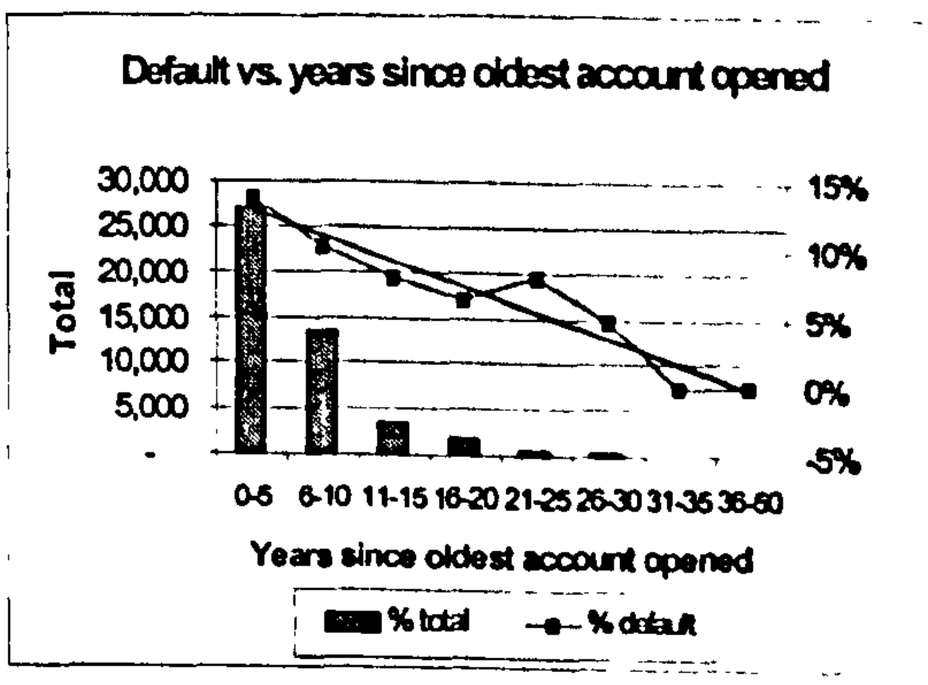
Figure 8.4.1.b.

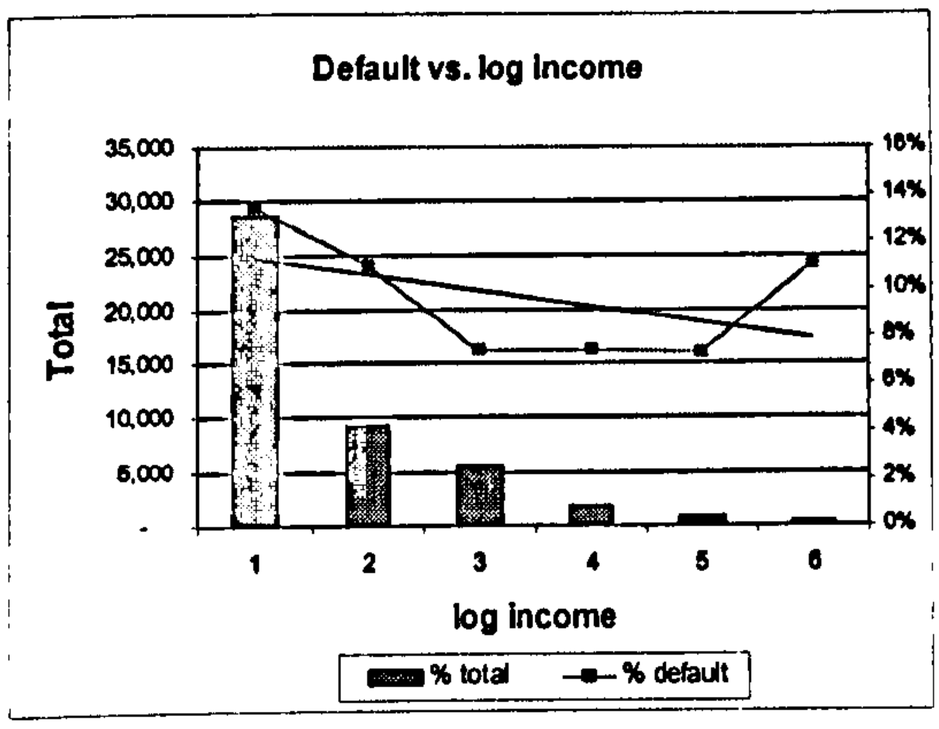
8.5 Appendix

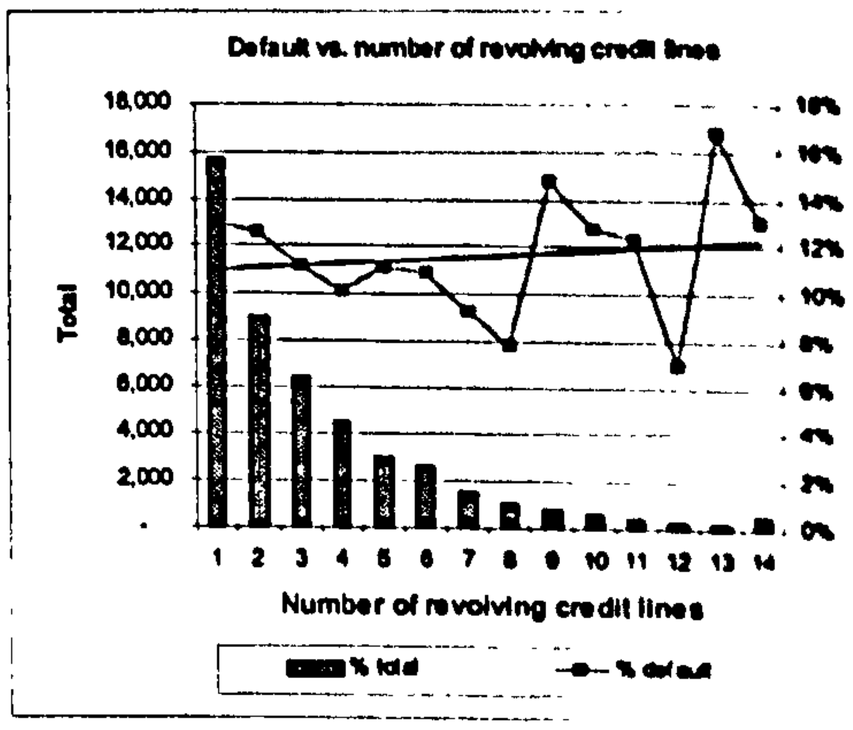

Default vs. economic dependants
Total
14,000
12,000
10,000
8,000
6,000
4,000
2,000
-
20%
15%
10%
5%
0%
1 2 3 4 5 6 7 8 9 10 11
economic dependants
% total
% default

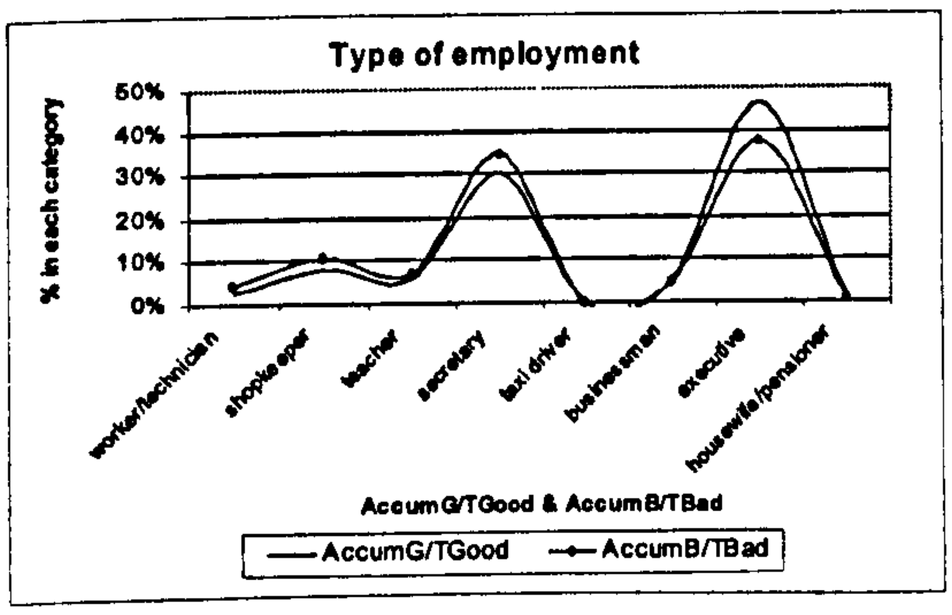

Default vs. number of satisfactory references
Total
20,000
15,000
10,000
5,000
-
20%
15%
10%
5%
0%
-5%
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of satisfactory references
% total
% default

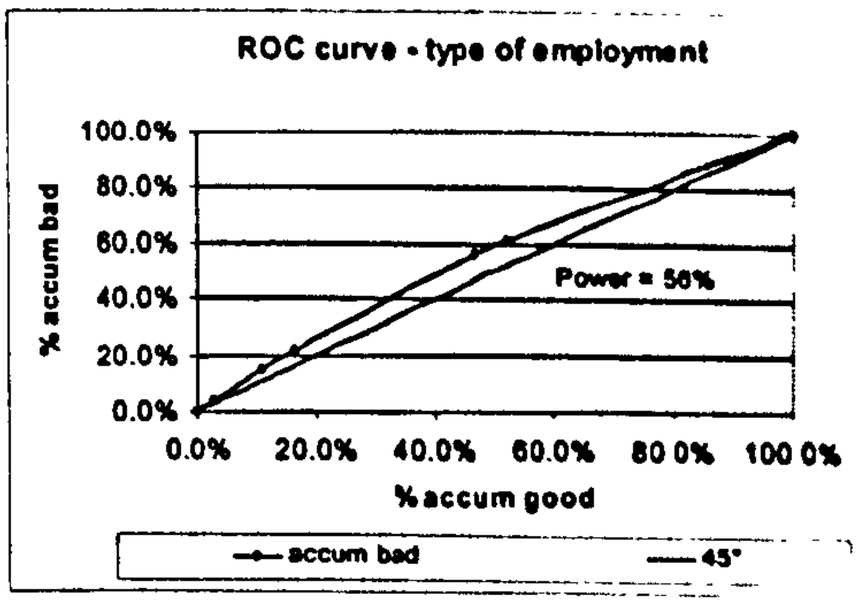

Default vs. age
Total
12,000
10,000
8,000
6,000
4,000
2,000
-
15%
10%
5%
0%
18-24 25-30 31-35 36-40 41-45 46-50 51-55 56-60 61-65 66-70
Age
% total
% default

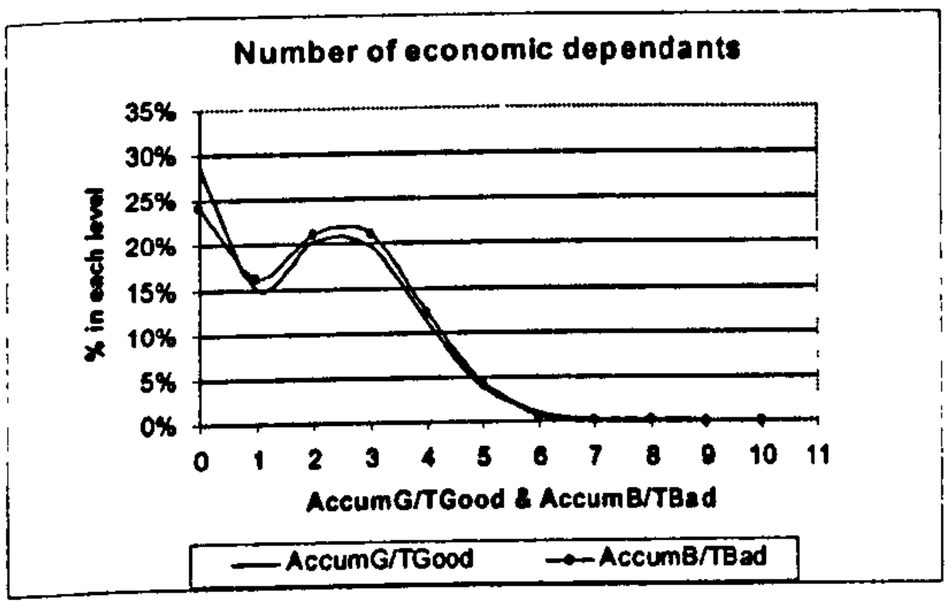

Default vs. years since oldest account opened
Total
30,000
25,000
20,000
15,000
10,000
5,000
-
15%
10%
5%
0%
-5%
0-5 6-10 11-15 16-20 21-25 26-30 31-35 36-60
Years since oldest account opened
% total
% default

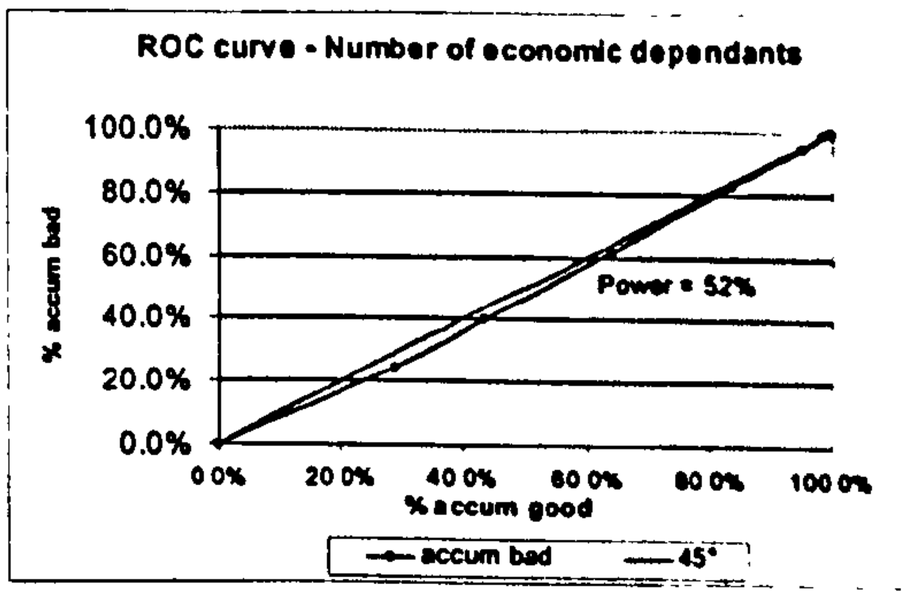

Default vs. log income
Total
log income
% total
% default

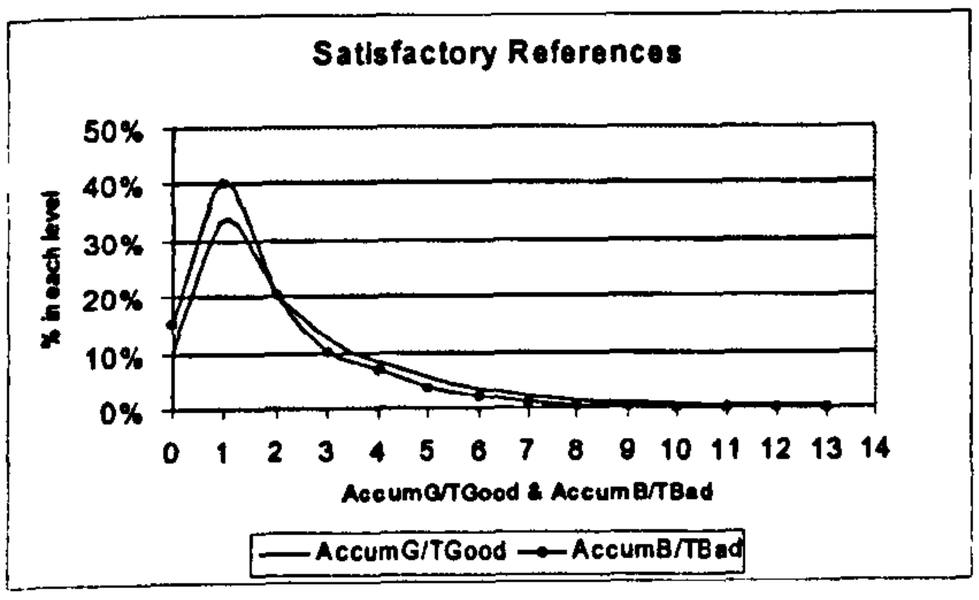

Default vs. number of revolving credit lines
Total
Number of revolving credit lines
% total
% default

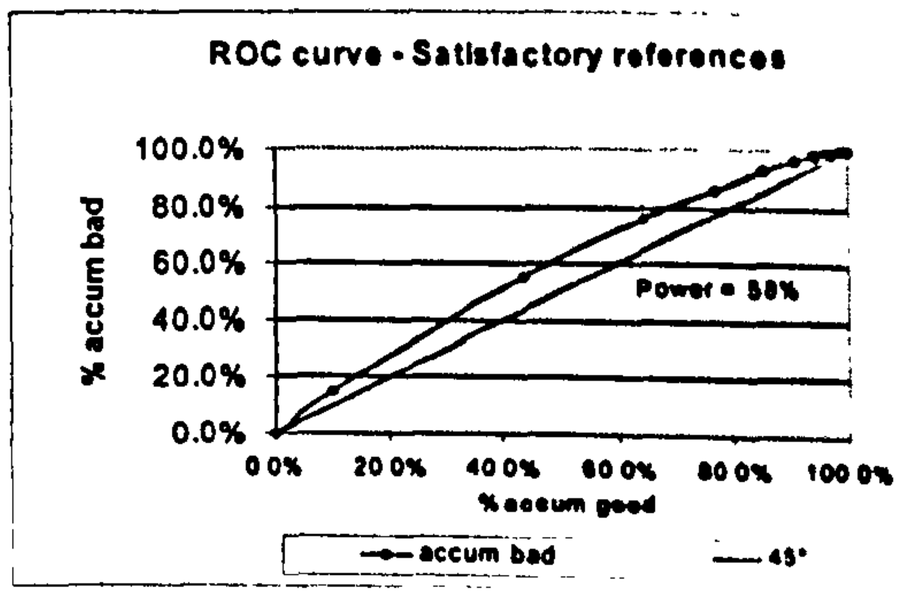

Type of employment
% in each category
worker/technician
shopkeeper
teacher
secretary
taxi driver
businessman
executive
housewife/pensioner
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad

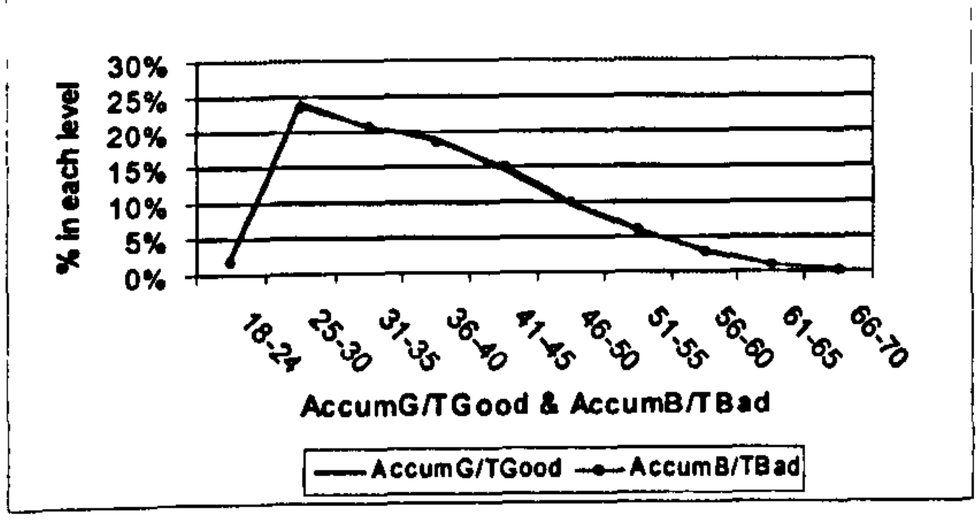

ROC curve - type of employment
% accum bad
Power = 56%
% accum good
accum bad
45°

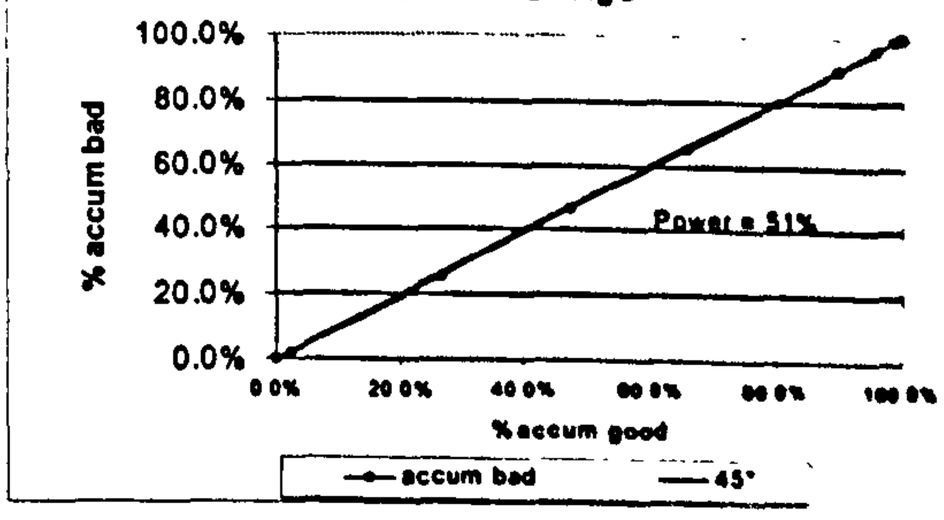

Number of economic dependants
% in each level
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad

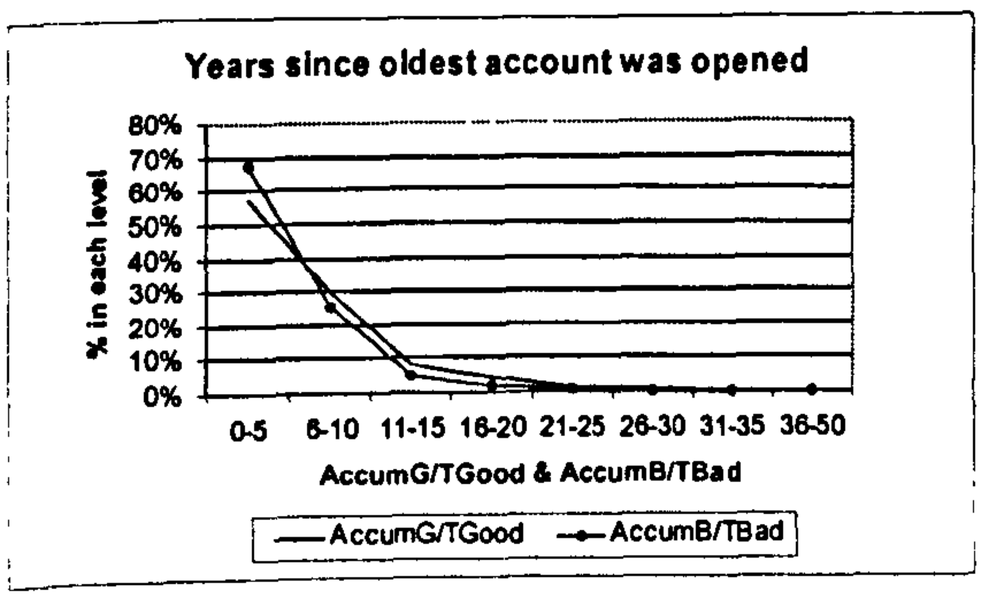

ROC curve - Number of economic dependants
% accum bad
Power = 52%
% accum good
accum bad
45°

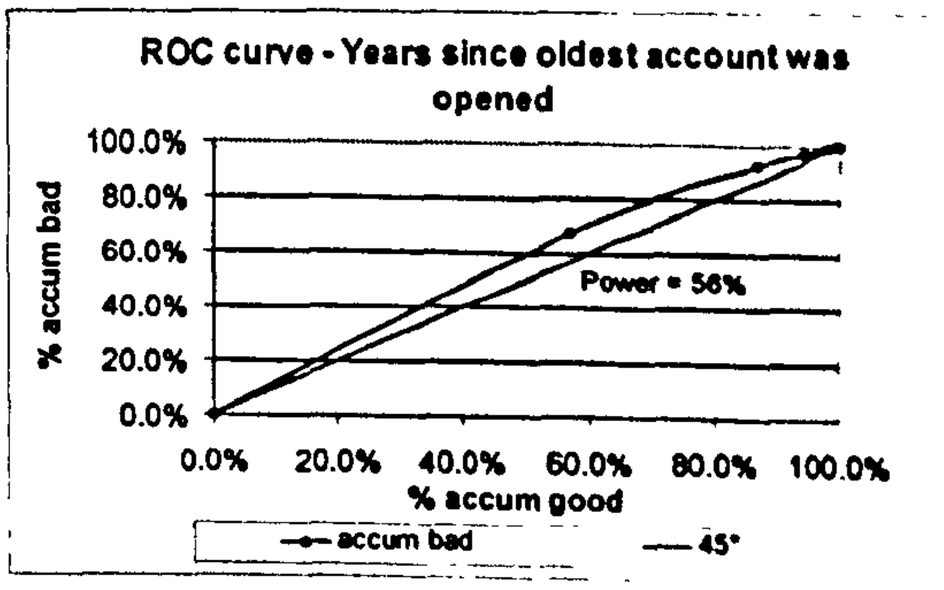

Satisfactory References
50%
40%
30%
20%
10%
0%
% in each level
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad

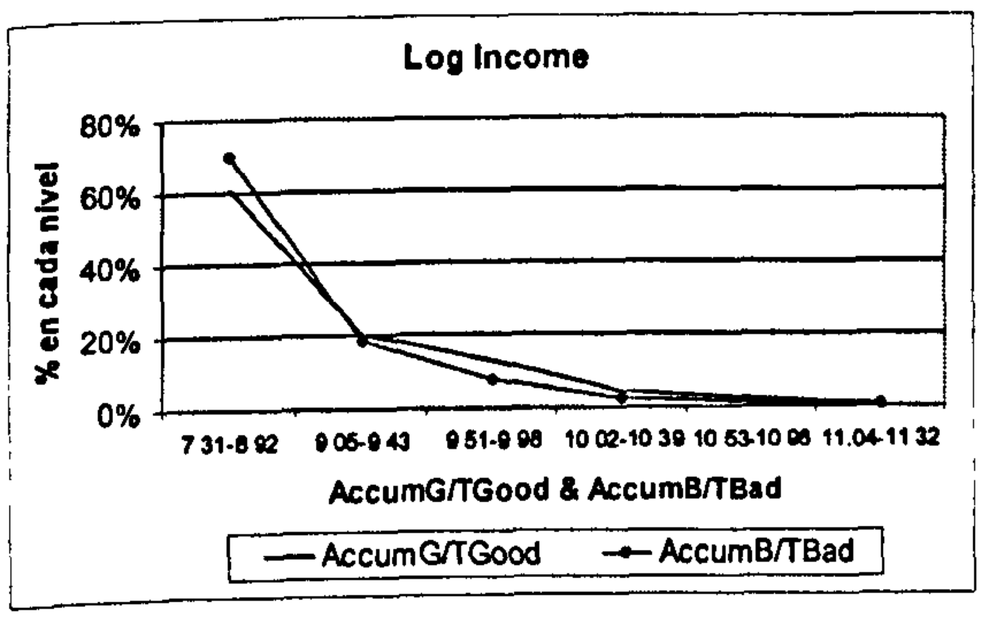

ROC curve - Satisfactory references
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
% accum bad
Power = 58%
0 0%
20 0%
40 0%
60 0%
80 0%
100 0%
% accum good
accum bad
45°

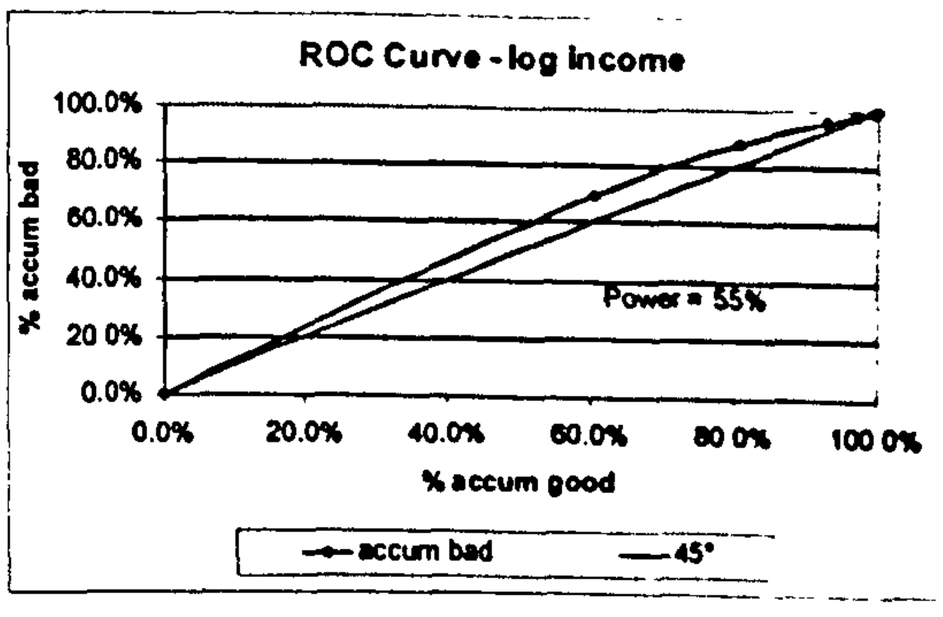

Age
30%
25%
20%
15%
10%
5%
0%
% in each level
18-24
25-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad

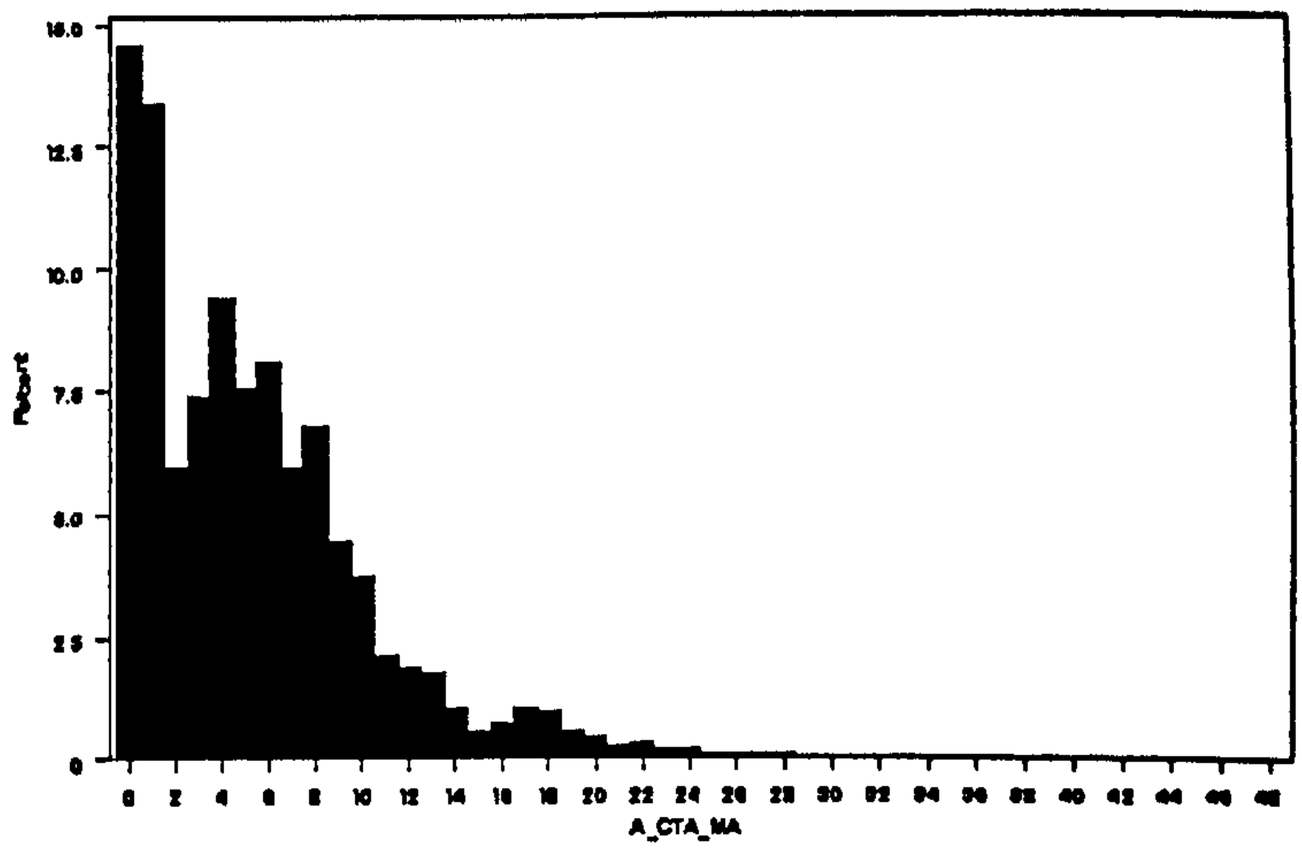

ROC curve - Age
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
% accum bad
Power = 51%
% accum good
accum bad
45°


Years since oldest account was opened
80%
70%
60%
50%
40%
30%
20%
10%
0%
% in each level
0-5
6-10
11-15
16-20
21-25
26-30
31-35
36-50
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad

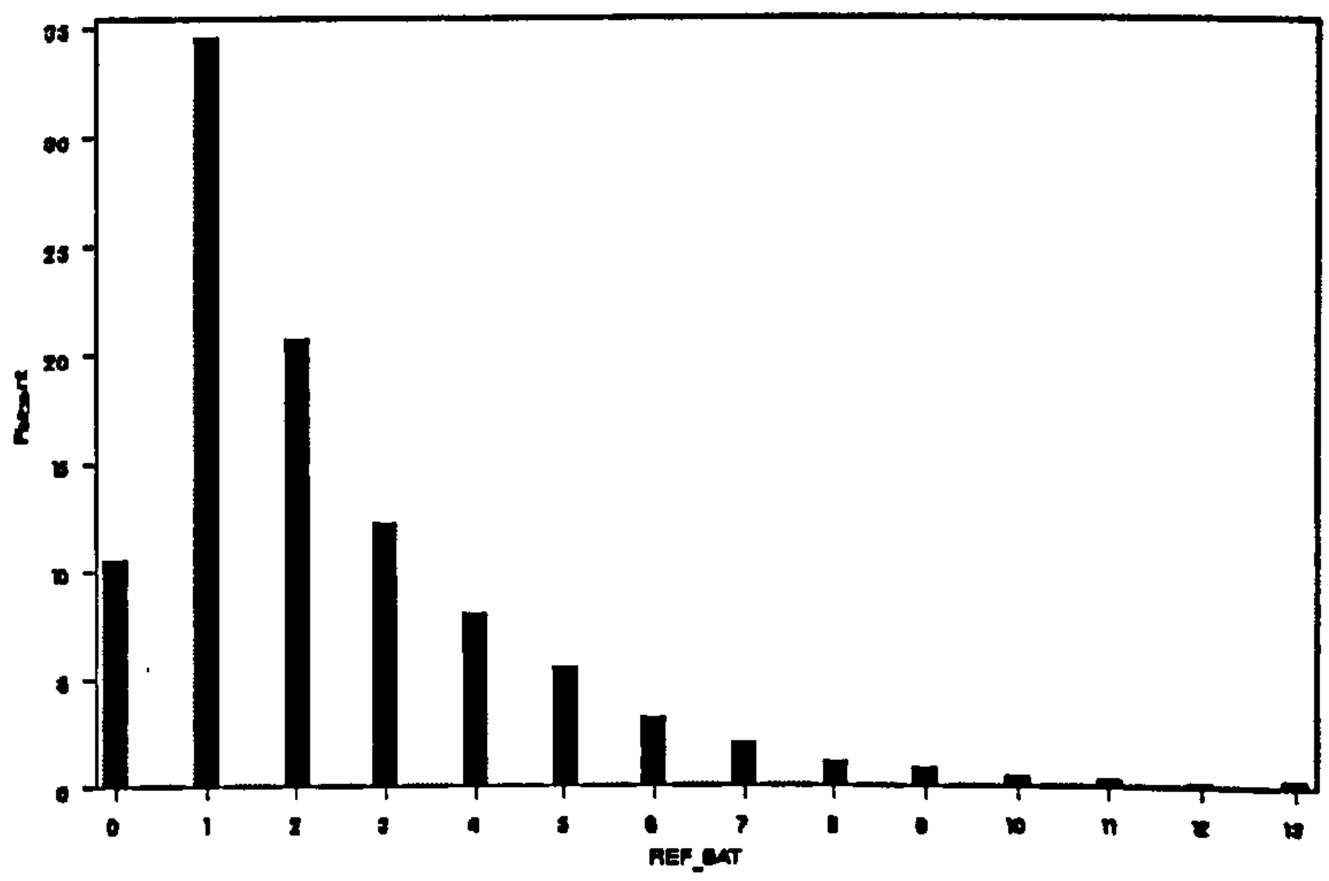

ROC curve - Years since oldest account was opened
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
% accum bad
Power = 56%
0.0%
20.0%
40.0%
60.0%
80.0%
100.0%
% accum good
accum bad
45°


Log Income
80%
60%
40%
20%
0%
% en cada nivel
7 31-8 92
9 05-9 43
9 51-9 98
10 02-10 39
10 53-10 98
11.04-11 32
AccumG/TGood & AccumB/TBad
AccumG/TGood
AccumB/TBad


ROC Curve - log income
100.0%
80.0%
60.0%
40.0%
20 0%
0.0%
% accum bad
Power = 55%
0.0%
20.0%
40.0%
60.0%
80 0%
100 0%
% accum good
accum bad
45°

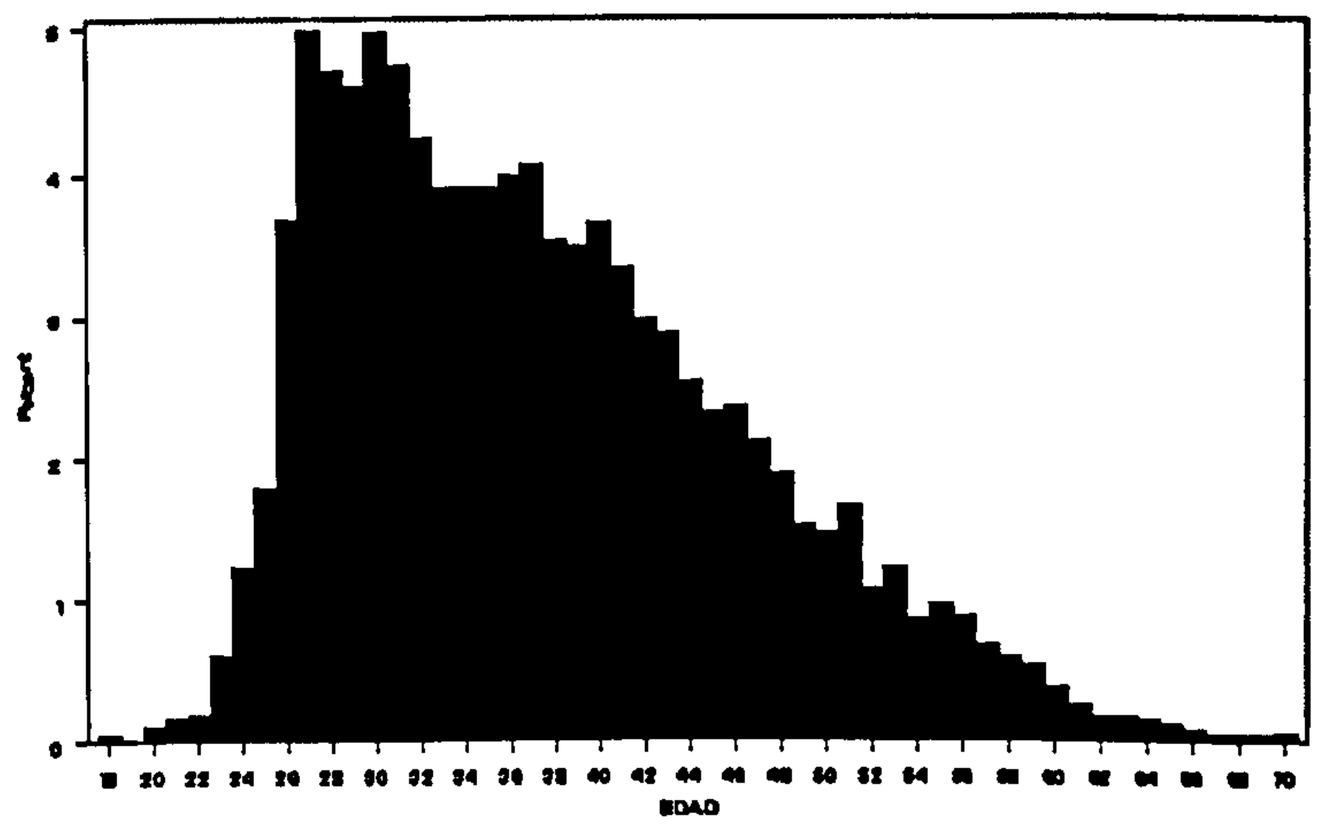


Histogram: Years since oldest account opened

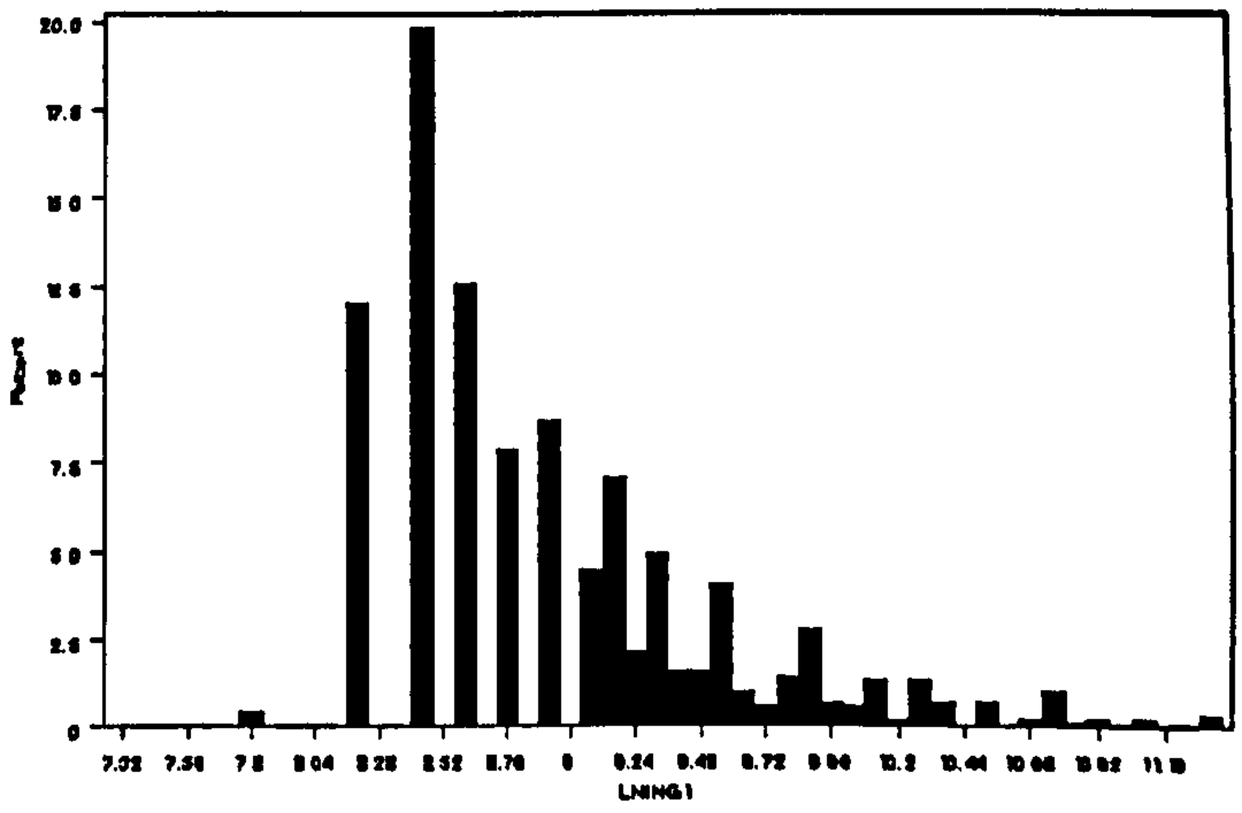


Histogram: Satisfactory References

Histogram: Economic dependants



Histogram: Age



Histogram: Natural logarithm of income

Chapter 9

Conclusions

Even though conclusions are given in each of the Chapters containing examples, a summary of these results is worth reviewing. The first evident conclusion is that the present method provides a useful technique when the underlying process generating the data is not known or if it is known to be largely biased with respect to a logistic model, which is the usual assumption. The behaviour of the non-parametric estimates resulted to be very satisfactory, even under simulations using a logistic underlying generation process. That is, the differences between the non-parametric and the logistic regression estimates were not significant for the logistic underlying setting. When the underlying setting was not logistic, the non-parametric estimates performed slightly better that logistic regression ones. The procedure also proved to be useful for applications other than credit scoring, such as quantal bioassay and medical screening.

Many aspects of this problem have yet to be explored in more detail. These include

the generalisations and transformation of covariates, which were topics that were not considered in depth. Also, convergence issues involved in the numerical approximations were in some cases difficult to deal with. This is another aspect that could be studied further. Finally, the case study presented represents a simplification of what could be done in a real life situation. This analysis depended on the amount of information that was available. Model builders in financial institutions will be able to apply this procedure in a more comprehensive manner if they deem it useful to their needs.

Bibliography

[1] Apostol, T. M (1967). *Calculus, One-variable Calculus, with an Introduction to Linear Algebra.* Massachusetts: Blaisdell Publishing Company.

[2] Beale, R. and Jackson, T. (1990) *Neural Computing - an Introduction.* IOP Publishing Ltd, Bristol.

[3] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Wadsworth International.

[4] Chatfield, C. and Collins, A. J. (1995). *Introduction to Multivariate Analysis.* London: Chapman & Hall.

[5] Cochran, W. G (1977). *Sampling Techniques.* New York: John Wiley & Sons.

[6] Cleveland, W. S. (1979) Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, Vol. 74, 829 – 836.

[7] Collet, D. (1992). *Modelling Binary Data*. London: Chapman & Hall.

[8] Copas, J. B. (1983). Plotting p against x. *The Journal of the Royal Statistical Society*, Vol. 32, No. 1, 25 – 31.

[9] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. New York: John Wiley & Sons, 2nd edition.

[10] Eguchi, S. and Copas, J.B.(2002). A class of logistic-type discriminant functions. *Biometrika*, Vol. 89, No. 1, 1 – 22.

[11] Fan, J., Heckman, N. E., and Wand, M. P. (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, Vol. 90, No. 429, 141 – 150.

[12] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7.

[13] Hand, D. J. (1992) Statistical methods in diagnosis. *Statistical Methods in Medical Research* 1, 49 – 67.

[14] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman & Hall.

[15] Henley, W. E. (1994). *Statistical aspects of credit scoring.* PhD Thesis. Department of Statistics, Faculty of Mathematics and Computing, The Open University.

[16] Hernández Arellano, F. M. (2003). *Cálculo de Probabilidades.* Sociedad Matemática Mexicana.

[17] Loftsgarden, D.O. and Queensberry, C.P. (1965) A nonparametric estimate of a multivariate density function. *Annals of Mathematics and Statistics* 36, 1049 – 1051.

[18] McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society (Series B)* 42, 109 – 142.

[19] McCullagh, P. and Nelder, J.A. (1983). *Generalised Linear Models.* London: Chapman & Hall.

[20] Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill.

[21] Müller, H-G and Schmitt, T. (1988). *Kernel and Probit Estimates in Quantal Bioassay.* Journal of the American Statistical Association, Vol. 83, No. 403, 750 – 759.

[22] Pack, S. E. and Morgan, B. J. T. (1990) A mixture model for interval-censored time-to-response quantal assay. *Biometrics*, Vol. 46, No. 3, 749 – 757.

[23] Sen, P. K. and Singer J. M. (1993). *Large Sample Methods in Statistics.* London: Chapman & Hall.

[24] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability 26, Chapman & Hall.

[25] Stone, C.J. (1977) Nonparametric regression and its applications (with discussion). *Annals of Statistics* 5, 595 – 645.

[26] Taylor, J. M. G. (1995) Semiparametric estimation in failure time mixture models. *Biometrics*, Vol. 51, No. 3, 899 – 907.

[27] Terrell, G.R. and Scott, D.W. (1992) Variable kernel density estimation. *The Annals of Statistics* 20 (3), 1236 – 1265.

[28] Tsodikov, A. D. et al (1995) Discrete stategies of cancer post-treatment surveillance estimation and optimization problems. *Biometrics*, Vol. 51, No. 2, 437 – 447.

[29] Venables, W. N. and Ripley, B. D (1994). *Modern Applied Statistics with S-plus.* New York: Springer-Verlag.